Adaptive Forward-Backward Greedy Algorithm for Learning Sparse Representations

Tong Zhang, *Member, IEEE*

***Abstract*—Given a large number of basis functions that can be potentially more than the number of samples, we consider the problem of learning a sparse target function that can be expressed as a linear combination of a small number of these basis functions. We are interested in two closely related themes:**

- **feature selection, or identifying the basis functions with nonzero coefficients;**
- **estimation accuracy, or reconstructing the target function from noisy observations.**

Two heuristics that are widely used in practice are forward and backward greedy algorithms. First, we show that neither idea is adequate. Second, we propose a novel combination that is based on the forward greedy algorithm but takes backward steps adaptively whenever beneficial. For least squares regression, we develop strong theoretical results for the new procedure showing that it can effectively solve this problem under some assumptions. Experimental results support our theory.

***Index Terms*—Estimation theory, feature selection, greedy algorithms, sparse recovery, statistical learning.**

I. Introduction

CONSIDER a set of input vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in R^d$, with corresponding desired output variables $y_1, \ldots, y_n$. The task of supervised learning is to estimate the functional relationship $y \approx f(\mathbf{x})$ between the input $\mathbf{x}$ and the output variable y from the training examples $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$. The quality of prediction is often measured through a loss function $\phi(f(\mathbf{x}), y)$. In this paper, we consider linear prediction model $f(\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x}$. A commonly used estimation method is empirical risk minimization

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in R^d} \sum_{i=1}^{n} \phi(\boldsymbol{\beta}^\top \mathbf{x}_i, y_i). \tag{1}$$

Note that in this paper, we are mainly interested in the least squares problem where $\phi(\boldsymbol{\beta}^\top \mathbf{x}_i, y_i) = (\boldsymbol{\beta}^\top \mathbf{x}_i - y_i)^2$.

In modern machine learning applications, one is typically interested in the scenario that $d \gg n$. That is, there are many more features than the number of samples. In this case, a direct application of (1) is inadequate because the solution of $\hat{\boldsymbol{\beta}}$ may not be unique (which is often referred to as *ill-posed* in the numerical computation literature). Statistically, the solution $\hat{\boldsymbol{\beta}}$ overfits the data. The standard remedy for this problem is to impose a regularization condition of $\boldsymbol{\beta}$ to obtain a *well-posed* problem. For computational reasons, one often employs a convex regularization condition which leads to a convex optimization problem of the following form:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in R^d} \sum_{i=1}^{n} \phi(\boldsymbol{\beta}^\top \mathbf{x}_i, y_i) + \lambda g(\boldsymbol{\beta}) \tag{2}$$

where $\lambda > 0$ is a tuning parameter, and $g(\boldsymbol{\beta})$ is a regularization condition, such as $g(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_p^p$.

One view of this additional regularization condition is that it constrains the target function space, which we assume can be approximated by some $\bar{\boldsymbol{\beta}}$ with small ℓ_p-norm $||\bar{\boldsymbol{\beta}}||_p$. An important target constraint is sparsity, which corresponds to the (nonconvex) L_0 regularization, where we let $||\bar{\boldsymbol{\beta}}||_0 = |\{j : \bar{\boldsymbol{\beta}}_j \neq 0\}| = k$. If we know the sparsity parameter k, a good learning method is L_0 regularization

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in R^d} \frac{1}{n} \sum_{i=1}^{n} \phi(\boldsymbol{\beta}^\top \mathbf{x}_i, y_i) \quad \text{subject to} ||\boldsymbol{\beta}||_0 \leq k. \tag{3}$$

If k is not known, then one may regard k as a tuning parameter, which can be selected through cross-validation. Sparse learning is an essential topic in machine learning, which has attracted considerable interests recently. Generally speaking, one is interested in two closely related themes:

- feature selection, or identifying the basis functions with nonzero coefficients;
- estimation accuracy, or reconstructing the target function from noisy observations.

If we can solve the first problem, that is, if we can perform feature selection well, then we can also solve the second problem. This is because after feature selection, we only need to perform empirical risk minimization (1) with the selected features. However, it is possible to obtain good prediction accuracy without solving the feature selection problem.

This paper focuses on the situation that approximate feature selection is possible. Under this scenario, we obtain results both on feature selection accuracy and on prediction accuracy (through oracle inequalities). Our main assumption is the *restricted isometry condition* (RIC) of [7], which we shall also refer to as the *sparse eigenvalue condition* in this paper. This condition says that any small number (at the order of the desired sparsity level) of features are not highly correlated. In fact, if a small number of features are correlated, then it is impossible to achieve accurate feature selection because a sparse target

Manuscript received October 16, 2008; revised October 05, 2010; accepted December 31, 2010. Date of current version June 22, 2011. The author was supported in part by the following grants: AFOSR-10097389, NSA-AMS 081024, NSF DMS-1007527, and NSF IIS-1016061.

The author is with the Statistics Department, Rutgers University, Piscataway NJ 08854 USA (e-mail: tzhang@stat.rutgers.edu).

Communicated by A. Krzyzak, Associate Editor for Pattern Recognition, Statistical Learning, and Inference.

Digital Object Identifier 10.1109/TIT.2011.2146690

may be represented using more than one set of sparse features. Therefore, the effectiveness of any feature selection algorithm requires such a condition.

While L_0 regularization in (3) is the most obvious formulation to solve the feature selection problem if the target function can be approximated by a sparse $\bar{\boldsymbol{\beta}}$, a fundamental difficulty with this method is the computational cost, because the number of subsets of $\{1,\ldots,d\}$ of cardinality k (corresponding to the nonzero components of $\boldsymbol{\beta}$) is exponential in k. There are no efficient algorithms to solve (3) in the general case.

Due to the computational difficult, in practice, there are several standard methods for learning sparse representations by solving approximations of (3). Their effectiveness has been recently analyzed under various assumptions.

- L_1-regularization (Lasso): the idea is to replace the L_0 regularization in (3) by L_1 regularization

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in R^d} \frac{1}{n}\sum_{i=1}^{n} \phi(\boldsymbol{\beta}^\top \mathbf{x}_i, y_i) \quad \text{subject to } \|\boldsymbol{\beta}\|_1 \le s,$$

  or equivalently, solving (2) with $p = 1$. This is the closest convex approximation to (3). It is known that L_1 regularization often leads to sparse solutions. Its performance has been studied recently. For example, if the target is truly sparse, then it was shown in [19], [29] that under some restricted conditions referred to as *irrepresentable conditions*, L_1 regularization solves the feature selection problem. However, such conditions are much stronger than RIC considered here. The prediction performance of L_1 regularization has been considered in [16], [3], [5]. Performance bounds can also be obtained when the target function is only approximately sparse (e.g., [27], [6]). Despite its popularity, there are several problems with L_1 regularization: first, the sparsity (L_0 complexity) is only implicitly controlled through L_1 approximation, which means that desirable feature selection property can only be achieved under relatively strong assumptions; second, in order to obtain very sparse solution, one has to use a large regularization parameter λ in (2) that leads to suboptimal prediction accuracy because the L_1 penalty not only shrinks irrelevant features to zero, but also shrinks relevant features to zero. A sub-optimal remedy is to threshold the resulting coefficients as suggested in [27] and to use two stage procedures. However, this approach requires additional tuning parameters, making the resulting procedures more complex and less robust.
- Forward greedy algorithm, which we will describe in details in Section II. The method has been widely used by practitioners. For least squares regression, this method is referred to as *matching pursuit* [18] in the signal processing community (also see [15], [1], [2], [9]). In machine learning, the method is often known as boosting [4], and similar ideas have been explored in Bayesian network learning [8]. The sparse regression performance of forward greedy algorithm has been analyzed in [23], [11] without considering stochastic noise, while its feature selection performance with stochastic noise has recently been studied in [26]. It was shown that the irrepresentable condition of [29] for L_1 regularization is also necessary for the greedy algorithm to effectively select features.
- Backward greedy algorithm, which we will describe in details in Section II. Although this method is widely used by practitioners, there isn't much theoretical analysis in the literature when $n \ll d$. The reason will be discussed in Section II. When $n \gg d$, backward greedy may be successful under some assumptions [10].

We shall point out that if we are only interested in prediction performance instead of feature selection, then the problem of learning sparse representation is also related to learning a sparse target function under many irrelevant features, which has long been studied in the online learning literature. In particular, exponentiated gradient descent methods such as Winnow are also effective [17]. However, this class of methods do not lead to sparse solutions. More recently, sparse online learning has attracted significant attention, and various work showed that this is possible to achieve through online L_1 regularization [12], [24]. However, such analysis only implies an oracle inequality with $O(1/\sqrt{n})$ convergence rate when the performance of online L_1 regression is compared to that of an arbitrary coefficient vector $\bar{\boldsymbol{\beta}}$. However, this performance guarantee is suboptimal in the sparse regression setting under sparse eigenvalue assumptions. For example, the oracle inequality in this paper would imply a bound of the form of no worse than $O(\|\bar{\boldsymbol{\beta}}\|_0 \ln d/n)$, which converges at a faster $O(1/n)$ rate (see Theorem 3.1). It remains an open problem to prove faster convergence rates for these recently proposed online algorithms in the high dimensional sparse regression setting.

In the batch learning setting, there has been considerable interest in learning sparse representations, and multiple algorithms have been proposed to solve the problem, satisfactory theoretical understanding (mainly for L_1 regularization) has only appeared very recently. In this paper, we are particularly interested in greedy algorithms because they have been widely used but the effectiveness has not been well analyzed. Moreover, they do not suffer from some shortcomings of L_1 regularization which we have pointed out earlier.

As we shall explain later, neither the standard forward greedy idea nor the standard backward greedy idea is adequate for our purpose. However, the flaws of these methods can be fixed by a simple combination of the two ideas. This leads to a novel adaptive forward-backward greedy algorithm which we present in Section III. The general idea works for all loss functions. For least squares loss, we obtain strong theoretical results showing that the method can solve the feature selection problem under moderate conditions.

For clarity, this paper only considers fixed design. In such case, the expectation $\mathbf{E}\mathbf{y}$ is always conditioned on the design matrix X. To simplify notations in our description, we will replace the optimization problem in (1) and (3) with a more general formulation. Instead of working with n input data vectors $\mathbf{x}_i \in R^d$, we work with d feature vectors $\mathbf{f}_j \in R^n$ ($j = 1,\ldots,d$), and $\mathbf{y} \in R^n$. Each $\mathbf{f}_j$ corresponds to the jth feature component of $\mathbf{x}_i$ for $i = 1,\ldots,n$. That is, $\mathbf{f}_{j,i} = \mathbf{x}_{i,j}$. Using

Input: $\mathbf{f}_1, \ldots, \mathbf{f}_d, \mathbf{y} \in R^n$ and $\epsilon > 0$
Output: $F^{(k)}$ and $\beta^{(k)}$
let $F^{(0)} = \emptyset$ and $\beta^{(0)} = 0$
for $k = 1, 2, \ldots$
  let $i^{(k)} = \arg\min_i \min_\alpha Q(\beta^{(k-1)} + \alpha \mathbf{e}_i)$
  let $F^{(k)} = \{i^{(k)}\} \cup F^{(k-1)}$
  let $\beta^{(k)} = \hat{\beta}(F^{(k)})$
  if $(Q(\beta^{(k-1)}) - Q(\beta^{(k)}) \leq \epsilon)$ **break**
end

Fig. 1. Forward Greedy Algorithm.

this notation, we can generally rewrite (3) with the problem of the following form:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in R^d} Q(\boldsymbol{\beta}) \quad \text{subject to } \|\boldsymbol{\beta}\|_0 \leq k$$

where $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d] \in R^d$ is the coefficient vector, and Q is defined below in (4).

In the following, we also let $\mathbf{e}_j \in R^d$ be the vector of zeros, except for the j-component which is one. Throughout the paper, we consider only the least squares loss

$$Q(\boldsymbol{\beta}) = \frac{1}{n} \left\| \mathbf{y} - \sum_{j=1}^{d} \boldsymbol{\beta}_j \mathbf{f}_j \right\|_2^2 = \frac{1}{n} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 \qquad (4)$$

where $\mathbf{y}^\top = [y_1, \ldots, y_n] \in R^n$, and we let $X = [\mathbf{f}_1, \ldots, \mathbf{f}_d]$ be the $n \times d$ data matrix.

For convenience, we also introduce the following notations.

Definition 1: Define $\text{supp}(\boldsymbol{\beta}) = \{j : \boldsymbol{\beta}_j \neq 0\}$ as the set of nonzero coefficients of a vector $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d] \in R^d$. Given $\mathbf{g} \in R^n$ and $F \subset \{1, \ldots, d\}$, let

$$\hat{\boldsymbol{\beta}}(F, \mathbf{g}) = \arg\min_{\boldsymbol{\beta} \in R^d} \frac{1}{n} \|X\boldsymbol{\beta} - \mathbf{g}\|_2^2 \text{ subject to } \text{supp}(\boldsymbol{\beta}) \subset F$$

and let $\hat{\boldsymbol{\beta}}(F) = \hat{\boldsymbol{\beta}}(F, \mathbf{y})$ be the solution of the least squares problem using feature set F.

Note that from the definition, $X\hat{\boldsymbol{\beta}}(F, \mathbf{g})$ is simply the projection of $\mathbf{g}$ to the subspace spanned by $\{\mathbf{f}_j : j \in F\}$.

II. Forward and Backward Greedy Algorithms

The forward greedy algorithm has been widely used in applications. It can be used to improve an arbitrary prediction method or select relevant features. In the former context, it is often referred to as boosting, and in the latter context, forward feature selection. Although a number of variations exist, they all share the basic form of greedily picking an additional feature at every step to aggressively reduce the squared error. The intention is to make most significant progress at each step in order to achieve sparsity. In this regard, the method can be considered as an approximation algorithm for solving (3). An example algorithm is presented in Fig. 1. This particular algorithm performs a full optimization using the selected basis function at each step, and is often referred to as orthogonal matching pursuit or OMP. This per-step optimization is important in our analysis.

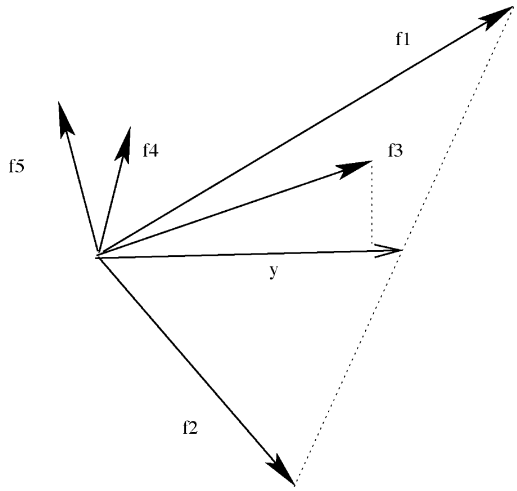

Fig. 2. Failure of Forward Greedy Algorithm.

A major flaw of this method is that it can never correct mistakes made in earlier steps. As an illustration, we consider the situation plotted in Fig. 2 with least squares regression. In the figure, $\mathbf{y}$ can be expressed as a linear combination of $\mathbf{f}_1$ and $\mathbf{f}_2$ but $\mathbf{f}_3$ is closer to $\mathbf{y}$. Therefore, using the forward greedy algorithm, we will find $\mathbf{f}_3$ first, then $\mathbf{f}_1$ and $\mathbf{f}_2$. At this point, we have already found all good features as $\mathbf{y}$ can be expressed by $\mathbf{f}_1$ and $\mathbf{f}_2$, but we are not able to remove $\mathbf{f}_3$ selected in the first step.

The above argument implies that forward greedy method is inadequate for feature selection. The method only works when small subsets of the basis functions $\{\mathbf{f}_j\}$ are near orthogonal. For example, see [23], [11] for analysis of greedy algorithm under such assumptions without statistical noise.[1] Its feature selection performance with stochastic noise has been recently studied in [26]. In general, Fig. 2 shows that even when the variables are not completely correlated (which is the case we consider in this paper), forward greedy algorithm will make errors that are not corrected later on. In fact, results in [23], [26] showed that in addition to the sparse eigenvalue condition, a stronger irrepresentable condition (also see [29]) is necessary for forward greedy algorithm to be successful.

For feature selection, the main problem of forward greedy algorithm is the lack of ability to correct errors made in earlier steps. In order to remedy the problem, the so-called backward greedy algorithm has been widely used by practitioners. The idea is to train a full model with all the features, and greedily remove one feature (with the smallest increase of squared error) at a time. The basic algorithm can be described in Fig. 3.

Although at the first sight, backward greedy method appears to be a reasonable idea that addresses the problem of forward greedy algorithm, it is computationally very costly because it starts with a full model with all features. Moreover, there are no theoretical results showing that this procedure is effective. In fact, under our setting, the method may only work when $d \ll n$

[1]Although the title in [11] claimed a treatment of noise, their definition of noise is not random, and thus different from ours. In our terminology, their noise only means that the target function is approximately sparse, which we also handle. It is different from stochastic noise considered in this paper.

Input: $\mathbf{f}_1, \ldots, \mathbf{f}_d, \mathbf{y} \in R^n$
Output: $F^{(k)}$ and $\beta^{(k)}$
let $F^{(d)} = \{1, \ldots, d\}$
for $k = d, d-1, \ldots$
  let $\beta^{(k)} = \hat{\beta}(F^{(k)})$
  let $j^{(k)} = \arg\min_{j \in F^{(k)}} Q(\hat{\beta}(F^{(k)} - \{j\}))$ $(*)$
  let $F^{(k-1)} = F^{(k)} - \{j^{(k)}\}$
end

Fig. 3. Backward Greedy Algorithm.

(see, for example, [10]), which is not the case we are interested in. In the case $d \gg n$, during the first step, $\boldsymbol{\beta}^{(d)}$ can immediately overfit the data with perfect prediction. Moreover, removing any feature does not help: that is, $\boldsymbol{\beta}^{(d-1)}$ still completely overfits the data no matter which feature (either relevant or irrelevant) is removed. Therefore, the method will completely fail when $d \gg n$, which explains why there is no theoretical result for this method.

It should be pointed out that the fundamental problem of backward greedy is that we cannot start with an overfitted model. To this end, one may replace $(*)$ in Fig. 3 by a solution procedure that does not overfit, for example, via L_1 regularization. Backward greedy combined with L_1 regularization is potentially beneficial because we can more effectively control the sparsity of the resulting L_1 regularization solution. However, such a procedure will be computationally costly, and its benefit is unknown. In this paper, we propose an alternative solution by combining the strength of both forward and backward greedy methods while avoiding their shortcomings.

III. Adaptive Forward-Backward Greedy Algorithm

As we have pointed out earlier, the main strength of forward greedy algorithm is that it always works with a sparse solution explicitly, and thus computationally efficient. Moreover, it does not significantly overfit the data due to the explicit sparsity. However, a major problem is its inability to correct any error made by the algorithm. On the other hand, backward greedy steps can potentially correct such an error, but need to start with a good model that does not completely overfit the data—it can only correct errors with a small amount of overfitting. Therefore, a combination of the two can solve the fundamental flaws of both methods. However, a key design issue is how to implement a backward greedy strategy that is provably effective. Some heuristics exist in the literature, although without any effectiveness proof. For example, the standard heuristics, described in [14] and implemented in SAS, includes another threshold ϵ' in addition to ϵ: a feature is deleted if the squared error increase by performing the deletion is no more than ϵ'. Unfortunately we cannot provide an effectiveness proof for this heuristics: if the threshold ϵ' is too small, then it cannot delete any spurious features introduced in the forward steps; if it is too large, then one cannot make progress because good features are also deleted. In practice it can be hard to pick a good ϵ', and even the best choice may be ineffective.

This paper takes a more principled approach, where we specifically design a forward-backward greedy procedure with *adaptive* backward steps that are carried out automatically.

Input: $\mathbf{f}_1, \ldots, \mathbf{f}_d, \mathbf{y} \in R^n$ and $\epsilon > 0$
Output: $F^{(k)}$ and $\beta^{(k)}$
let $\nu = 0.5$ (it can also be any number in $(0, 1)$)
let $F^{(0)} = \emptyset$ and $\beta^{(0)} = 0$
let $k = 0$
while (**true**)
  // forward step
  let $i^{(k)} = \arg\min_i \min_\alpha Q(\beta^{(k)} + \alpha \mathbf{e}_i)$
  let $F^{(k+1)} = \{i^{(k)}\} \cup F^{(k)}$
  let $\beta^{(k+1)} = \hat{\beta}(F^{(k+1)})$
  let $\delta^{(k+1)} = Q(\beta^{(k)}) - Q(\beta^{(k+1)})$
  if $(\delta^{(k+1)} \leq \epsilon)$
    break
  end
  let $k = k + 1$
  // backward step
  while (**true**)
    let $j^{(k)} = \arg\min_{j \in F^{(k)}} Q(\beta^{(k)} - \beta_j^{(k)} \mathbf{e}_j)$
    let $d^- = [Q(\beta^{(k)} - \beta_{j^{(k)}}^{(k)} \mathbf{e}_{j^{(k)}}) - Q(\beta^{(k)})]$
    let $d^+ = \delta^{(k)}$
    if $(d^- > \nu d^+)$
      break
    end
    let $k = k - 1$
    let $F^{(k)} = F^{(k+1)} - \{j^{(k+1)}\}$
    let $\beta^{(k)} = \hat{\beta}(F^{(k)})$
  end
end

Fig. 4. FoBa: Adaptive Forward-Backward Greedy Algorithm.

The procedure has provably good performance and fixes the drawbacks of forward greedy algorithm illustrated in Fig. 2. There are two main considerations in our approach.

- We want to take reasonably aggressive backward steps to remove any errors caused by earlier forward steps, and to avoid maintaining a large number of basis functions.
- We want to take backward step *adaptively* and make sure that any backward greedy step does not erase the gain made in the forward steps. This ensures that we are always making progress.

Our algorithm, which we refer to as *FoBa*, is listed in Fig. 4. It is designed to balance the above two aspects. Note that we only take a backward step when the squared error increase (d^-) is no more than half of the squared error decrease in the earlier corresponding forward step (d^+). This implies that if we take ℓ forward steps, then no matter how many backward steps are performed, the squared error is decreased by at least an amount of $\ell\epsilon/2$. It follows that if $Q(\boldsymbol{\beta}) \geq 0$ for all $\boldsymbol{\beta} \in R^d$, then the algorithm terminates after no more than $2Q(0)/\epsilon$ steps. This means that the procedure is computationally efficient.

Proposition 3.1: When the FoBa procedure terminates in Fig. 4, the total number of forward steps is no more than $1 + 2Q(0)/\epsilon$. Moreover, the total number of backward steps is no more than the total number of forward steps.

According to our theoretical results (e.g., Theorem 3.1 below), in order to achieve optimal performance, when the sample size n increases, we should decrease the stopping parameter ϵ as $\epsilon = O(\sigma^2 \ln d/n)$ (where σ^2 is the variance

of the noise). This implies that in general, the algorithm takes more steps to reach optimal performance when the sample size n increases. However, the number of steps presented in Proposition 3.1 is not tight. For example, in the scenario that FoBa can select all features correctly, it is possible for FoBa to take a constant number of steps to achieve correct feature selection independent of the sample size. Since sharp numerical convergence bound for FoBa is not the focus of the current paper, we shall not include a more refined analysis here.

Note that the claim of Proposition 3.1 (as well as the later theoretical analysis) still holds if we employ a more aggressive backward strategy as follows: set $d^- = d^+ = 0$ at the beginning of the backward step, and update the quantities as $d^- = d^- + [Q(\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k)}_{j^{(k)}}\mathbf{e}_{j^{(k)}}) - Q(\boldsymbol{\beta}^{(k)})]$ and $d^+ = d^+ + \delta^{(k)}$. That is, d^- and d^+ are cumulative changes of squared error.

Now, consider an application of FoBa to the example in Fig. 2. Again, in the first three forward steps, we will be able to pick $\mathbf{f}_3$, followed by $\mathbf{f}_1$ and $\mathbf{f}_2$. After the third step, since we are able to express $\mathbf{y}$ using $\mathbf{f}_1$ and $\mathbf{f}_2$ only, by removing $\mathbf{f}_3$ in the backward step, we do not increase the squared error. Therefore, at this stage, we are able to successfully remove the incorrect basis $\mathbf{f}_3$ while keeping the good features $\mathbf{f}_1$ and $\mathbf{f}_2$. This simple illustration demonstrates the effectiveness of FoBa.

In the following, we will formally characterize this intuitive example, and prove results for the effectiveness of FoBa for feature selection as well as prediction accuracy under the condition that the target is either truly sparse or approximately sparse. Since the situation in Fig. 2 is covered by our analysis, one cannot derive a similar result for the forward greedy algorithm. That is, the condition of our results do not exclude forward steps from making errors. Therefore, it is essential to include backward steps in our theoretical analysis.

We introduce the following definition, which characterizes how linearly independent small subsets of $\{\mathbf{f}_j\}$ of size k are. For $k \ll n$, the number $\rho(k)$ defined below can be bounded away from zero even when $d \gg n$. For example, for random basis functions $\mathbf{f}_j$, we may take $\ln d = O(n/k)$ and still have $\rho(k)$ to be bounded away from zero. This quantity is the smallest eigenvalue of the $k \times k$ principal submatrices of the $d \times d$ design matrix $X^\top X = [\mathbf{f}_i^\top \mathbf{f}_j]_{i,j=1,\ldots,d}$, and has appeared in recent analysis of L_1 regularization methods such as in [5], [25], [27], etc. It was first introduced in [7], and was referred to as the restricted isometry condition. In this paper, we shall also call it *sparse eigenvalue condition*. This condition is the least restrictive condition when compared to other conditions in the literature such as the irrepresentable condition [29] or the mutual coherence condition [11] (also see discussions in [3], [25], [27]).

Definition 3.1: Define for all integer $k \geq 1$

$$\rho(k) = \inf\left\{ \frac{1}{n}\|X\boldsymbol{\beta}\|_2^2 \Big/ \|\boldsymbol{\beta}\|_2^2 : \|\boldsymbol{\beta}\|_0 \leq k \right\}.$$

Assumption 3.1: Assume that the basis functions are normalized such that $\frac{1}{n}\|\mathbf{f}_j\|_2^2 = 1$ for all $j = 1, \ldots, d$, and assume that $\{y_i\}_{i=1,\ldots,n}$ are independent (but not necessarily identically distributed) sub-Gaussians: there exists $\sigma \geq 0$ such that $\forall i$ and $\forall t \in R$

$$\mathbf{E}_{y_i} e^{t(y_i - \mathbf{E}y_i)} \leq e^{\sigma^2 t^2/2}.$$

Both Gaussian and bounded random variables are sub-Gaussian using the above definition. For example, we have the following well-known Hoeffding's inequality.

Proposition 3.2: If a random variable $\xi \in [a, b]$, then $\mathbf{E}_\xi e^{t(\xi - \mathbf{E}\xi)} \leq e^{(b-a)^2 t^2/8}$. If a random variable is Gaussian: $\xi \sim N(0, \sigma^2)$, then $\mathbf{E}_\xi e^{t\xi} \leq e^{\sigma^2 t^2/2}$.

We will present a number of theoretical results for the FoBa algorithm. For convenience in our analysis and theoretical statements, we shall state all results with an explicitly chosen stopping parameter $\epsilon > 0$. In real applications, one can always run the algorithm with a smaller ϵ and use cross-validation to determine the optimal stopping point.

Our first result is an oracle inequality for the estimated coefficient vector $\boldsymbol{\beta}^{(k)}$, described in the following theorem.

Theorem 3.1: Consider the FoBa algorithm in Fig. 4 where Assumption 3.1 holds. Consider any approximate target vector $\bar{\boldsymbol{\beta}} \in R^d$ with $\bar{F} = \mathrm{supp}(\bar{\boldsymbol{\beta}})$, and $\bar{k} = |\bar{F}|$. Let $s \leq d$ be an integer that satisfies $32(\bar{k} + 1) \leq (0.8s - 2\bar{k})\rho(s)^2$. Assume that we set $\epsilon \geq 108\rho(s)^{-1}\sigma^2 \ln(16d)/n$ in the FoBa algorithm. Then for all $\eta \in (0, 1)$, with probability larger than $1 - \eta$, the following statements hold.

If FoBa terminates at some $k \leq s - \bar{k}$, then

$$\begin{aligned}&\left\|X\boldsymbol{\beta}^{(k)} - \mathbf{E}\mathbf{y}\right\|_2^2 - \|X\bar{\boldsymbol{\beta}} - \mathbf{E}\mathbf{y}\|_2^2 \\ &\leq 74\sigma^2\bar{k} + 27\sigma^2 \ln(2e/\eta) + 18n\epsilon\rho(s)^{-1}\Delta k(\epsilon, s)\end{aligned}$$

where $\Delta k(\epsilon, s) = |\{j \in \bar{F} : \bar{\beta}_j^2 < 32\epsilon\rho(s)^{-2}\}|$.

If FoBa terminates at some $k \geq s - \bar{k}$, then for all $k_0 \in [0.8s - \bar{k}, s - \bar{k}]$, at the start of the forward step when FoBa reaches $k = k_0$ for the first time, we have

$$\begin{aligned}&\|X\boldsymbol{\beta}^{(k)} - \mathbf{E}\mathbf{y}\|_2^2 \\ &\leq \|X\bar{\boldsymbol{\beta}} - \mathbf{E}\mathbf{y}\|_2^2 - 0.7n\bar{k}\epsilon + 10.8\ln(2e/\eta)\sigma^2.\end{aligned}$$

In order for the theorem to apply, we require a condition $32(\bar{k} + 1) \leq (0.8s - 2\bar{k})\rho(s)^2$, which can be satisfied as long as $\rho(s)$ is lower bounded by a constant at some sparsity level $s = O(\bar{k})$. We have already mentioned earlier that this type of condition (some times referred to as RIC [7]), while varying in details from one analysis to another (for example, see [3], [21]), is standard for analyzing algorithms for high dimensional sparsity problems such as Lasso. The details are not important for the purpose of this paper. Therefore, in our discussion, we shall simply assume it is satisfied at some s that is $O(\bar{k})$, and focus on the consequences of our theoretical results when this condition holds.

If we choose $\epsilon = O(\sigma^2 \ln d/n)$, then the worst case oracle inequality is of the form

$$\begin{aligned}\|X\boldsymbol{\beta}^{(k)} - \mathbf{E}\mathbf{y}\|_2^2 \leq \|X\bar{\boldsymbol{\beta}} - \mathbf{E}\mathbf{y}\|_2^2 \\ + O\left(\bar{k} + \ln(1/\eta) + \Delta k(\epsilon, s)\ln d\right)\sigma^2 \quad (5)\end{aligned}$$

where $\Delta k(\epsilon, s)$ is the number of nonzero coefficients in $\bar{\boldsymbol{\beta}}$ smaller than $O(\sigma \ln d/\sqrt{n})$. This bound applies when FoBa stops at a relatively sparse solution with $k \leq s - \bar{k}$. It is the sharpest possible bound, as the term $O(\bar{k} + \ln(1/\eta))$ is the parametric rate. The dimensional dependent term $\ln d$ involves the number $\Delta k(\epsilon, s)$ that is no more than $\bar{k}$, but can be much smaller; in fact $\Delta k(\epsilon, s) = 0$ when we choose a vector $\bar{\boldsymbol{\beta}}$ such that all its nonzero coefficients are larger than the order $\sigma \ln d/\sqrt{n}$.

The oracle inequality (5) is useful when FoBa stops at a solution $\boldsymbol{\beta}^{(k)}$ that is sparse. This happens when $\mathbf{E}\mathbf{y} \approx X\bar{\boldsymbol{\beta}}$ for some sparse approximate target vector $\bar{\boldsymbol{\beta}}$, as we will show in Theorem 3.2. The reason we can get this sharp oracle inequality for FoBa is because FoBa has good feature selection property, where most coefficients larger than the order $\sigma \ln d/\sqrt{n}$ can be correctly identified by the algorithm. This point will become more clear in Theorem 3.2 and Theorem 3.3. While the sparse regression setting is what we are mostly interested in this paper, it is worth noting that Theorem 3.1 can be applied even when $\mathbf{E}\mathbf{y}$ cannot be approximated very well by $X\bar{\boldsymbol{\beta}}$ with any sparse coefficient vector $\bar{\boldsymbol{\beta}}$. In fact, in such case, FoBa may not stop at $k < s - \bar{k}$. However, this is the relatively easy case because the last displayed inequality in Theorem 3.1 shows that along the FoBa path, the first time FoBa reaches any $k \in [0.8s - \bar{k}, s - \bar{k}]$ we can expect an even better oracle inequality

$$\left\|X\boldsymbol{\beta}^{(k)} - \mathbf{E}\mathbf{y}\right\|_2^2 \leq \|X\bar{\boldsymbol{\beta}} - \mathbf{E}\mathbf{y}\|_2^2 - n\bar{k}\epsilon/2 \tag{6}$$

as long as the chosen ϵ also satisfies a very mild condition

$$n\epsilon \geq \frac{54\sigma^2}{\bar{k}} \ln(2e/\eta).$$

This sharper bound might appear surprising at first. However, the reason that we can do better than any sparse $\bar{\boldsymbol{\beta}}$ is only because if $\mathbf{E}\mathbf{y}$ cannot be approximated by $X\bar{\boldsymbol{\beta}}$ with $\|\bar{\boldsymbol{\beta}}\|_0 = \bar{k}$, then we may find another coefficient vector with s nonzero-coefficients that can approximate $\mathbf{E}\mathbf{y}$ much better than $\bar{\boldsymbol{\beta}}$ does. Since the stopping criterion of FoBa guarantees that each forward greedy step reduces $Q(\cdot)$ by at least ϵ (which in the theorem is set to a number significantly larger than the noise level), it is therefore possible for FoBa to achieve better performance than that of $\bar{\boldsymbol{\beta}}$ as long as k is sufficiently large (but not too large to cause significant overfitting). The result shows that there is a wide range of values $k \in [0.8s - \bar{k}, s - \bar{k}]$ that can achieve this performance gain, and hence, it is relatively easy to identify such a k by cross validation. Nevertheless, this situation is of little interest in the current paper because if in a practical application, the target cannot be well approximated by a sparse vector, then sparse regression is not the best model for the problem. Therefore, although (6) may still be interesting theoretically, it does not have much value for practical purposes. Therefore, we will focus on the true sparsity case in the following analysis.

Theorem 3.1 shows that if FoBa reaches $k \geq s - \bar{k}$, then we can obtain an oracle inequality with performance better than that of any competitive $\bar{\boldsymbol{\beta}}$ at sparsity level $\|\bar{\boldsymbol{\beta}}\|_0 = \bar{k}$. The previous discussion indicates that this can only happen when any sparse $\bar{\boldsymbol{\beta}}$ does not approximate the true target well, which is of little interest in this paper. The following theorem shows that if the target is approximately sparse, that is, $\mathbf{E}\mathbf{y} \approx X\bar{\boldsymbol{\beta}}$, then FoBa will terminate at some k that is not much larger than $\bar{k}$, when we choose ϵ appropriately. In such case, the oracle inequality given in (5) is more meaningful, and it represents the best possible performance one can achieve in sparse regression. Again we would like to emphasize that the fundamental reason behind this sharp oracle inequality for the FoBa algorithm is due to the ability of FoBa to select quality features, which is presented below in Theorem 3.3. These results give better insights into the FoBa algorithm when the target is sparse.

Theorem 3.2: Consider the FoBa algorithm in Fig. 4 with a stopping parameter $\epsilon > 0$, where Assumption 3.1 holds. Consider any approximate target vector $\bar{\boldsymbol{\beta}} \in R^d$ with $\bar{F} = \mathrm{supp}(\bar{\boldsymbol{\beta}})$, and $\bar{k} = |\bar{F}|$. Let $s \leq d$ be an integer that satisfies $8(\bar{k} + 1) \leq (s - 2\bar{k})\rho(s)^2$. Then for all $\eta \in (0, 1)$, if

$$n\epsilon > \frac{2\rho(s)}{\bar{k}+1}\|X\bar{\boldsymbol{\beta}} - \mathbf{E}\mathbf{y}\|_2^2 + 5.4\sigma^2[8\rho(s)^{-1}\ln(16d) + \rho(s)(\bar{k}+1)^{-1}\ln(2e/\eta)]$$

then with probability larger than $1 - 2\eta$, FoBa terminates at $k < s - \bar{k}$.

Moreover, if $\epsilon \geq 108\rho(s)^{-1}\sigma^2 \ln(16d)/n$, then

$$\left\|X\boldsymbol{\beta}^{(k)} - \mathbf{E}\mathbf{y}\right\|_2^2 - \|X\bar{\boldsymbol{\beta}} - \mathbf{E}\mathbf{y}\|_2^2 \leq 74\sigma^2\bar{k} + 27\sigma^2\ln(2e/\eta) + 18n\epsilon\rho(s)^{-1}\Delta k(\epsilon, s)$$

where $\Delta k(\epsilon, s) = |\{j \in \bar{F} : \bar{\boldsymbol{\beta}}_j^2 < 32\epsilon\rho(s)^{-2}\}|$.

The above theorem shows that we may take

$$n\epsilon = O\left(\frac{2\rho(s)}{\bar{k}+1}\|X\bar{\boldsymbol{\beta}} - \mathbf{E}\mathbf{y}\|_2^2 + 5.4\sigma^2[8\rho(s)^{-1}\ln(16d) + \rho(s)(\bar{k}+1)^{-1}\ln(2e/\eta)]\right)$$

so that FoBa stops at $k < s - \bar{k}$. It implies an oracle inequality of the form

$$\left\|X\boldsymbol{\beta}^{(k)} - \mathbf{E}\mathbf{y}\right\|_2^2 \leq c\|X\bar{\boldsymbol{\beta}} - \mathbf{E}\mathbf{y}\|_2^2 + O(\bar{k} + \ln(1/\eta) + \ln(d)\Delta k(\epsilon, s))\sigma^2$$

with $c > 1$. Although on surface, this oracle inequality is slightly worse than (5), which has $c = 1$ (achieved at a potentially smaller ϵ), the difference is only due to some technical details in our analysis, as we shall note that Theorem 3.2 is mostly useful when

$$\|X\bar{\boldsymbol{\beta}} - \mathbf{E}\mathbf{y}\|_2^2 = O(\sigma^2\bar{k} + \sigma^2\ln(2e/\eta) + \sigma^2\ln(16d)\Delta k(\epsilon, s))$$

which means that $\mathbf{E}\mathbf{y}$ can be approximated very well by $X\bar{\boldsymbol{\beta}}$ with a sparse coefficient vector $\bar{\boldsymbol{\beta}}$. In this situation, the oracle inequality in Theorem 3.2 is comparable to that of Theorem 3.1. Note that if there is a sparse vector $\bar{\boldsymbol{\beta}}$ such that $\mathbf{E}\mathbf{y} = X\bar{\boldsymbol{\beta}}$, then Theorem 3.2 immediately implies that we can recover the true parameter $\bar{\boldsymbol{\beta}}$ using FoBa. For completeness, we include the following result as a direct consequence of Theorem 3.2.

Corollary 3.1: If the assumptions of Theorem 3.2 hold, and we further assume that $\mathbf{E}\mathbf{y} = X\bar{\boldsymbol{\beta}}$, then

$$\|\boldsymbol{\beta}^{(k)} - \bar{\boldsymbol{\beta}}\|_2^2 \leq 74\sigma^2\rho(s)^{-1}\bar{k} + 27\sigma^2\rho(s)^{-1}\ln(2e/\eta) + 18n\epsilon\rho(s)^{-2}\Delta k(\epsilon, s)$$

where $\Delta k(\epsilon, s) = |\{j \in \bar{F} : \bar{\boldsymbol{\beta}}_j^2 < 32\epsilon\rho(s)^{-2}\}|$.

If $\mathbf{E}\mathbf{y} \approx X\bar{\boldsymbol{\beta}}$ with a sparse coefficient vector $\bar{\boldsymbol{\beta}}$, then FoBa is able to correctly select feature coefficients that are significantly differentiable from the noise level. This feature selection property is the reason why we are able to obtain the sharp form of oracle inequalities in Theorem 3.1 and in Theorem 3.2. The feature selection quality of FoBa is presented in Theorem 3.3. For simplicity, we only consider the situation that $\mathbf{E}\mathbf{y} = X\bar{\boldsymbol{\beta}}$ (although similar to Theorem 3.2, we may allow a small approximation error).

Theorem 3.3: Consider the FoBa algorithm in Fig. 4 with some $\epsilon > 0$, where Assumption 3.1 holds. Assume that the true target vector is sparse: that is, there exists $\bar{\boldsymbol{\beta}} \in R^d$ such that $X\bar{\boldsymbol{\beta}} = \mathbf{E}\mathbf{y}$, with $\bar{F} = \mathrm{supp}(\bar{\boldsymbol{\beta}})$, and $\bar{k} = |\bar{F}|$. Let $s \leq d$ be an integer that satisfies $8(\bar{k}+1) \leq (s - 2\bar{k})\rho(s)^2$. Then for all $\eta \in (0, 1)$, if

$$n\epsilon > 5.4\sigma^2\rho(s)^{-1}[8\ln(16d) + 2\rho(s)^2\ln(2e/\eta)]$$

then with probability larger than $1 - 3\eta$, FoBa terminates at $k < s - \bar{k}$, and

$$\rho(s)^2\left|F^{(k)} - \bar{F}\right| \Big/ 32 \leq \left|\bar{F} - F^{(k)}\right| \leq 2\left|\{j \in \bar{F} : \bar{\boldsymbol{\beta}}_j^2 < 32\epsilon\rho(s)^{-2}\}\right|.$$

It is worth mentioning that in Theorem 3.2 and Theorem 3.3, the specific choice of s is only used to show that FoBa stops at $s - \bar{k}$, and this specific choice can be improved with a more refined analysis. Moreover, s can be replaced by any upper bound of $|\bar{F} \cup F^{(k)}|$ when FoBa stops without changing either the feature selection result or the oracle inequality when FoBa terminates.

The proofs of the above theorems can be found in Appendix A. The high level idea, which relies on properties of the FoBa procedure at the stopping time, is described there before the technical details. In particular, the proofs do not attempt to show that any particular backward step is effective. Instead, for feature selection, we can show that if a particular backward step deletes a good feature, then FoBa does not stop; FoBa does not stop if many good features in $\bar{F}$ are missing from $F^{(k)}$; the backward step will start to delete (either bad or good) features if $F^{(k)}$ contains too many features not in $\bar{F}$. By combining these properties, we can deduce that once the FoBa procedure stops, the feature set $F^{(k)}$ approximately equals $\bar{F}$. Since our analysis does not require the effectiveness of a specific backward step, our results do not hold for a simpler procedure that performs a sequence of forward steps, followed by a sequence of backward steps. Such a method is unreliable because good features can be deleted in the backward steps. This is why we call this procedure "adaptive", and our adaptive forward-backward approach does not suffer from the problem in the simpler method.

Let

$$\Delta k(\epsilon, s) = \left|\left\{j \in \bar{F} : \bar{\boldsymbol{\beta}}_j^2 < 32\epsilon\rho(s)^{-2}\right\}\right|$$

be the number of very small nonzero coefficients of $\bar{\boldsymbol{\beta}}$ that are at the noise level. Therefore, $\Delta k(\epsilon, s)$ provides a nature quantity of nonzero coefficients of $\bar{\boldsymbol{\beta}}$ that cannot be differentiated from zero in the presence of noise.

Theorem 3.3 shows that when FoBa stops, we have

$$\left|\bar{F} - F^{(k)}\right| = O(\Delta k(\epsilon, s)), \left|F^{(k)} - \bar{F}\right| = O(\Delta k(\epsilon, s))$$

which implies that the estimated feature set $F^{(k)}$ differs from $\bar{F}$ by no more than $O(\Delta k(\epsilon, s))$ elements. In particular, if $\Delta k(\epsilon, s) = 0$, then we can reliably recover the true feature $\bar{F}$ with large probability. This is the fundamental reason why we can obtain a sharp oracle inequality as in (5). Note that only the last term in (5) depends on the dimension d, which is small when $\Delta k(\epsilon, s)$ is small: this happens when only a small number of $j \in \bar{F}$ have small coefficients $|\bar{\boldsymbol{\beta}}_j|^2 \leq O(\epsilon)$. In the worst case, even when all coefficients are small (and thus reliable feature selection is impossible), the bound is still meaningful with $\Delta k(\epsilon, s) = \bar{k}$. It leads to a bound

$$\left\|X\boldsymbol{\beta}^{(k)} - \mathbf{E}\mathbf{y}\right\|_2^2 \leq \|X\bar{\boldsymbol{\beta}} - \mathbf{E}\mathbf{y}\|_2^2 + O\left(\ln(1/\eta) + \bar{k}\ln d\right)\sigma^2$$

that becomes similar to oracle inequality for the Lasso under appropriate sparse eigenvalue assumptions. The result in (5) shows that it is possible for FoBa to achieve better performance than that of Lasso when most features can be reliably identified (that is, when $\Delta k(\epsilon, s)$ is small). This theoretical benefit is real, and is confirmed with our experiments.

A useful application of Theorem 3.2 and Theorem 3.3 is when $\mathbf{E}\mathbf{y} = X\bar{\boldsymbol{\beta}}$, and $\bar{\boldsymbol{\beta}}$ has relatively large nonzero coefficients. That is, we can identify the correct set of features with large probability. This particular problem has drawn significant interests in recent years, such as [29], [27].

Corollary 3.2: Consider the FoBa algorithm in Fig. 4, where Assumption 3.1 holds. Consider a target $\bar{\boldsymbol{\beta}} \in R^d$ such that $\mathbf{E}\mathbf{y} = X\bar{\boldsymbol{\beta}}$. Let $s \leq d$ be an integer such that $8(\bar{k}+1) \leq (s-2\bar{k})\rho(s)^2$, and assume that $|\bar{\boldsymbol{\beta}}_j|^2 > 32\epsilon/\rho(s)^2$ for all $j \in \bar{F}$. Assume that for some $\eta \in (0, 1/3)$, we have

$$n\epsilon > \max[5.4\sigma^2\rho(s)^{-1}(8\ln(16d) + 2\rho(s)^2\ln(2e/\eta)) \quad 108\rho(s)^{-1}\sigma^2\ln(16d)].$$

Then with probability larger than $1 - 3\eta$: when the algorithm terminates, we have $\bar{F} = F^{(k)}$ and

$$\left\|X\left(\boldsymbol{\beta}^{(k)} - \bar{\boldsymbol{\beta}}\right)\right\|_2^2 \leq 74\sigma^2\bar{k} + 27\sigma^2\ln(2e/\eta).$$

The corollary says that we can identify the correct set of features $\bar{F}$ as long as the coefficients $\bar{\boldsymbol{\beta}}_j$ $(j \in \bar{F})$ are larger than the noise level at some $O(\sigma\sqrt{\ln d/n})$. Such a requirement is quite natural, and occurs in other work on the effectiveness of feature selection [29], [27]. In fact, if any nonzero coefficient is

below the noise level, then no algorithm can distinguish it from zero with large probability. That is, it is impossible to reliably perform feature selection due to the noise. Therefore, FoBa is near optimal in term of its ability to perform reliable feature selection, except for the constant hiding in $O(\cdot)$ (as well as its dependency on $\rho(s)$).

The result can be applied as long as eigenvalues of small $s \times s$ diagonal blocks of the design matrix $X^\top X$ are bounded away from zero (that is, the sparse eigenvalue condition holds). This is the situation under which the forward greedy step can make mistakes, but such mistakes can be corrected using FoBa. Because the conditions of the corollary do not prevent forward steps from making errors, the example described in Fig. 2 indicates that it is not possible to prove a similar result for the forward greedy algorithm. In fact, it was shown in [26] that the stronger irrepresentable condition of [29] is necessary for the forward greedy algorithm to be effective.

IV. FoBa Versus Lasso

Lasso can successfully select features under *irrepresentable conditions* of [29] (also see [23] which considered the noiseless case). It was shown that such a condition is necessary for feature selection using Lasso, when zero-threshold is used. It is known that the sparse eigenvalue condition considered in this paper is significantly weaker [25], [20], [3].

Although under the sparse eigenvalue condition, it is not possible to reliably select features using Lasso and zero-threshold, it was pointed out in [27] that it is possible to reliably select features using an appropriately chosen nonzero threshold (that is, a postprocessing step is used to remove features with coefficients smaller than a certain threshold). There are two problems for this approach. First, this requires tuning two parameters: one is the L_1 regularization parameter, and the other nonzero threshold. Second, even if one can tune the threshold parameter successfully, the result in [27] requires that the condition $\min_{j\in\bar{F}} |\bar{\boldsymbol{\beta}}_j|^2 \geq c\bar{k}\sigma^2 \ln(d/\eta)/n$ for some $c > 0$ under the sparse eigenvalue condition considered here (although the $\bar{k}$-dependence can be removed under stronger assumptions). This is due to the inclusion of L_1 regularization which introduces a "bias". In comparison, Theorem 3.3 only requires $\min_{j\in\mathrm{supp}(\bar{\boldsymbol{\beta}})} |\bar{\boldsymbol{\beta}}_j|^2 \geq c\sigma^2 \ln(d/\eta)/n$ for some $c > 0$. The difference can be significant when $\bar{k}$ is large.

There is no counterparts of Theorem 3.1, Theorem 3.2, and Theorem 3.3 for L_1 regularization under the sparse eigenvalue conditions. The closest analogy is a parameter estimation bound for a multistage procedure investigated in [28], which solves a nonconvex optimization problem by using multiple stages of convex relaxation.

Finally, we shall point out that the forward-backward greedy procedure is closely related to the path-following algorithm for solving Lasso, such as the LARS algorithm for solving Lasso in [13], where one starts with a very large (infinity) regularization parameter, which is gradually decreased. Similar to FoBa, LARS also has forward and backward steps. However, unlike FoBa, which tracks the insertion and deletion with unbiased least squares error, LARS tracks the path through L_1 penalized least squares error, by gradually decreasing the regularization parameter. Initially, to obtain a very sparse set of features, one has to set a very large L_1 regularizaton parameter, which causes a significant bias. The added bias implies that LARS deviates significant from subset selection at least initially. When the algorithm progresses, the regularization parameter is reduced, and thus, the extra L_1 bias becomes smaller. Since the theory of L_1 regularization requires a regularization parameter larger than the noise level, the resulting bias is not negligible. Similar to FoBa, some mistakes made in earlier steps can be potentially removed later on.

It follows from the above discussion that FoBa is related to the path-following view of L_1 regularization. However, unlike Lasso, FoBa does not introduces a bias. Instead, it generates the path by directly keeping track of the objective function. Therefore, FoBa is closer to subset selection than L_1 regularization, especially when a highly sparse solution is desired (as far as training error is concerned). This claim is confirmed by our experiments.

V. Experiments

We compare FoBa to forward-greedy and L_1-regularization on artificial and real data. They show that in practice, FoBa is closer to subset selection than the other two approaches, in the sense that FoBa achieves smaller training error given any sparsity level. In order to compare with Lasso, we use the LARS [13] package in R, which generates a path of actions for adding and deleting features, along the L_1 solution path. For example, a path of $\{1, 3, 5, -3, \ldots\}$ means that in the fist three steps, feature 1, 3, 5 are added; and the next step removes feature 3.

Using such a solution path, we can compare Lasso to Forward-greedy and FoBa under the same framework. Similar to the Lasso path, FoBa also generates a path with both addition and deletion operations, while forward-greedy algorithm only adds features without deletion. Our experiments compare the performance of the three algorithms using the corresponding feature addition/deletion paths. We are interested in features selected by the three algorithms at any sparsity level k, where k is the desired number of features presented in the final solution. Given a path, we can keep an active feature set by adding or deleting features along the path. For example, for path $\{1, 3, 5, -3\}$, we have two potential active feature sets of size $k = 2$: $\{1, 3\}$ (after two steps) and $\{1, 5\}$ (after four steps). We then define the k best features as the active feature set of size k with the smallest least squares error because this is the best approximation to subset selection (along the path generated by the algorithm). For the forward-greedy algorithm, this is just the first k features in the path. For FoBa, it is the last time when the active set contains k features, because the squared error is always reduced in later stages. For Lasso, it is also likely to be the last time when the active set contains k features—this is because it corresponds to the smallest regularization parameter (thus, smallest bias). However, to be safe, for Lasso, we compute the least squares solutions for all active feature sets of size k, and then select the one with the smallest squared error.

From the above discussion, we do not have to set ϵ explicitly in the FoBa procedure. Instead, we just generate a solution path which is five times as long as the maximum desired sparsity k, and then generate the best k features for any sparsity level using the above described procedure.

TABLE I
PERFORMANCE COMPARISON ON SIMULATION DATA AT SPARSITY LEVEL $k = 5$

	FoBa	Forward-greedy	L_1
least squares training error	0.093 ± 0.02	0.16 ± 0.089	0.25 ± 0.14
parameter estimation error	0.057 ± 0.2	0.52 ± 0.82	1.1 ± 1
feature selection error	0.76 ± 0.98	1.8 ± 1.1	3.2 ± 0.77

A. Simulation Data

Since for real data, we do not know the true feature set $\bar{F}$, simulation is needed to compare feature selection performance. We generate $n = 100$ data points of dimension $d = 500$. The target vector $\bar{\boldsymbol{\beta}}$ is truly sparse with $\bar{k} = 5$ nonzero coefficients generated uniformly from 0 to 10. The noise level is $\sigma^2 = 0.1$. The basis functions $\mathbf{f}_j$ are randomly generated with moderate correlation: that is, some basis functions are correlated to the basis functions spanning the true target. However, the correlation is relatively small, and thus does not violate the RIC condition. Note that if there is no correlation (i.e., $\mathbf{f}_j$ are independent random vectors), then all three methods work well (this is the well-known case considered in the compressed sensing literature). If the correlation among $\mathbf{f}_j$ is very strong, then it is not surprising that all three methods will fail. Therefore, in this experiment, we generate moderate correlation so that the performance of the three methods can be differentiated. Such moderate correlation does not violate the sparse eigenvalue condition in our analysis, but violates the more restrictive conditions for forward-greedy method and Lasso.

Table I shows the performance of the three methods, where we repeat the experiments 50 times, and report the average $\pm$ standard-deviation. We use the three methods to select five best features, using the procedure described above. We report three metrics. Training error is the squared error of the least squares solution with the selected five features. Parameter estimation error is the 2-norm of the estimated parameter (with the five features) minus the true parameter. Feature selection error is the number of incorrectly selected features. It is clear from the table that for this data, FoBa achieves significantly smaller training error than the other two methods, which implies that it is closest to subset selection. Moreover, the parameter estimation performance and feature selection performance are also better. Note that in this example, the noise level $\sigma/\sqrt{n}$ is relatively small compared to the range of nonzero coefficients of the ground truth, which ensures that feature selection can be performed relatively reliably. If feature selection becomes unreliable (that is, when we increase σ significantly), then L_1 regularization may achieve better performance because it is inherently more stable.

B. Real Data

Instead of listing results for many datasets, we consider two data sets that reflect typical behaviors of the algorithms. A careful analysis of the two datasets leads to better insights than a list of performance numbers for many data. The experiments show that FoBa does what it is designed to do well: that is, it gives a better approximation to subset selection than either forward-greedy or L_1 regularization. However, as well shall see, better sparsity does not always lead to better generalization on real data. This is because sparsity alone is not necessarily always the best complexity measure for real problems.

1) Boston Housing Data: The first dataset we consider is the *Boston Housing* data, which is the housing data for 506 census tracts of Boston from the 1970 census, available from the *UCI Machine Learning Database Repository*: *http://archive.ics.uci.edu/ml/*. Each census tract is a data-point, with 13 features (we add a constant offset one as the 14th feature), and the desired output is the housing price. In the experiment, we randomly partition the data into 50 training plus 456 test points. We perform the experiments 50 times, and for each sparsity level from 1 to 10, we report the average training and test squared error. The results are plotted in Fig. 5. From the results, we can see that FoBa achieves better training error for any given sparsity, which is consistent with the theory and the design goal of FoBa. Moreover, it achieves better test accuracy with small sparsity level (corresponding to a more sparse solution). With large sparsity level (corresponding to a less sparse solution), the test error increase more quickly with FoBa. This is because it searches a larger space by more aggressively mimicking subset selection, which makes it more prone to overfitting. However, at the best sparsity level of 3, FoBa achieves significantly better test error. Moreover, we can observe with small sparsity level (a more sparse solution), L_1 regularization performs poorly, due to the bias caused by using a large L_1-penalty.

For completeness, we also compare FoBa to the backward-greedy algorithm and the classical heuristic forward-backward greedy algorithm as implemented in SAS (see its description at the beginning of Section III). We still use the Boston Housing data, but plot the results separately, in order to avoid cluttering. As we have pointed out, there is no theory for the SAS version of forward-backward greedy algorithm. It is difficult to select an appropriate backward threshold ϵ': a too small value leads to few backward steps, and a too large value leads to overly aggressive deletion, and the procedure terminates very early. In this experiment, we pick a value of 10, because it is a reasonably large quantity that does not lead to an extremely quick termination of the procedure. The performance of the algorithms are reported in Fig. 6. From the results, we can see that backward greedy algorithm performs reasonably well on this problem. Note that for this data, $d \ll n$, which is the scenario that backward does not start with a completely overfitted full model. Still, it is inferior to FoBa at small sparsity level, which means that some degree of overfitting still occurs. Note that backward-greedy algorithm cannot be applied in our simulation data experiment, because $d \gg n$ which causes immediate overfitting. From the graph, we also see that FoBa is more effective than the SAS implementation of forward-backward greedy algorithm. The latter does not perform significant better than the forward-greedy algorithm with our choice of ϵ'. Unfortunately, using a larger backward threshold ϵ' will lead to an undesirable early termination of the

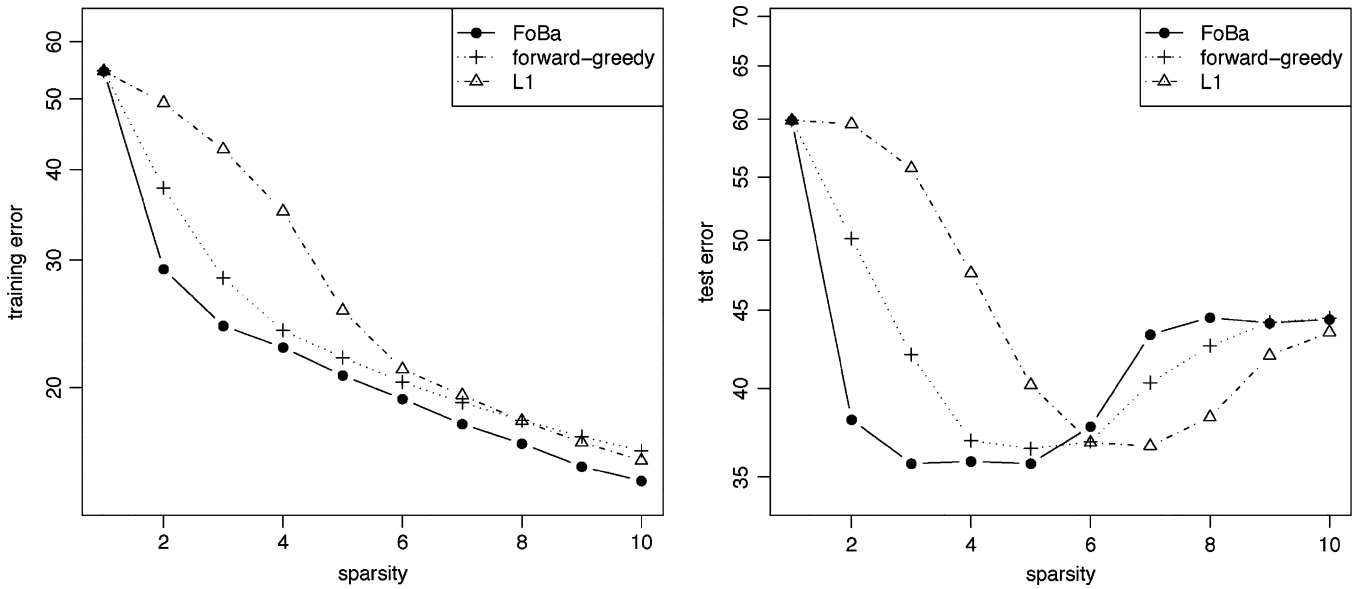

Fig. 5. Performance of the algorithms on *Boston Housing* data Left: average training squared error versus sparsity; Right: average test squared error versus sparsity.

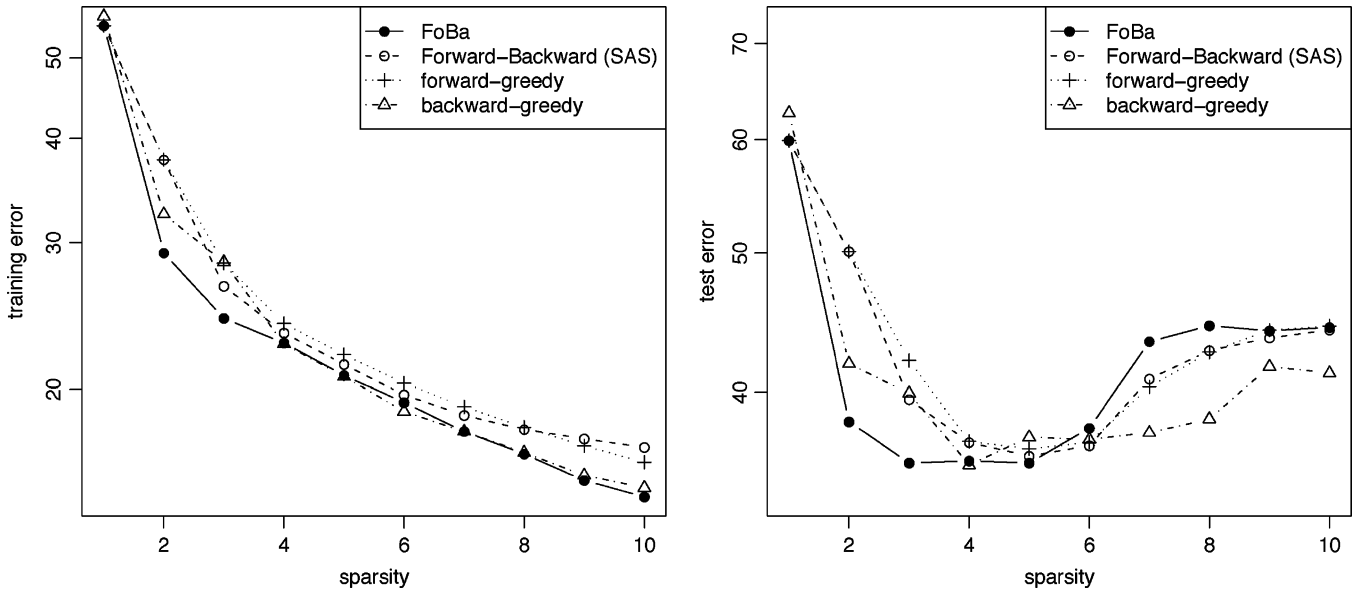

Fig. 6. Performance of greedy algorithms on *Boston Housing* data. Left: average training squared error versus sparsity; Right: average test squared error versus sparsity.

algorithm. This intuition is why the provably effective adaptive backward strategy introduced in this paper is superior.

2) Ionosphere Data: The second dataset we consider is the *Ionosphere* data, also available from the *UCI Machine Learning Database Repository*. It contains 351 data points, with 34 features (we add a constant offset one as the 35th feature), and the desired output is binary valued $\{0, 1\}$. Although this is a classification problem, we treat it as regression. In the experiment, we randomly partition the data into 50 training plus 301 test points. We perform the experiments 50 times, and for each sparsity from 1 to 10, we report the average training and test squared error. The results are plotted in Fig. 7. From the results, we can see that FoBa again achieves better training error for any given sparsity, which is consistent with the theory and the design goal of FoBa. However, it does not achieve better test accuracy. This suggests that sparsity alone is not the correct complexity measure for this data. Indeed by examining the results closely, we observe that the coefficients of Lasso solution tend to be much smaller (due to the added L_1 constraints) than those from FoBa or the forward-greedy algorithm (which do not favor small coefficients in their designs). From Fig. 7, we can see that even with smaller coefficients, Lasso achieves similar training error at small sparsity level. This means that Lasso effectively searches a smaller space, and thus more stable and less prone to overfitting. Therefore, for this dataset, the added prior knowledge of small coefficients (in addition to sparsity) in L_1 regularization

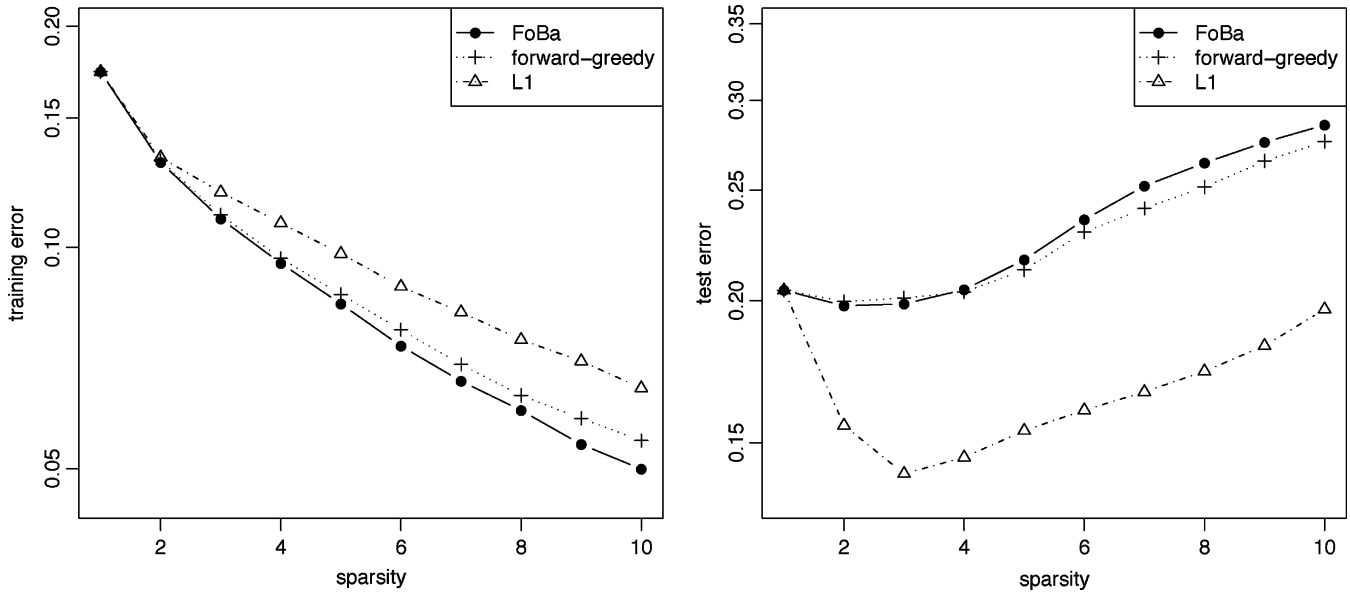

Fig. 7. Performance of the algorithms on *Ionosphere* data. Left: average training squared error versus sparsity; Right: average test squared error versus sparsity.

gives it an edge over the greedy approaches, which do not take the size of coefficients into consideration. For simplicity, we do not include a comparison with the backward greedy method.

VI. Discussion

This paper investigates the problem of learning sparse representations using greedy algorithms. We showed that neither forward greedy nor backward greedy algorithms are adequate by themselves. However, through a novel combination of the two ideas, we proved that an adaptive forward-back greedy algorithm, referred to as FoBa, can effectively solve the problem under reasonable conditions.

FoBa is designed to be a better approximation to subset selection. In fact, backward step naturally appears in solving L_0 regularized optimization problems. Consider the penalization version of (3)

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in R^d} \left[\frac{1}{n}\sum_{i=1}^{n} \phi(\boldsymbol{\beta}^\top \mathbf{x}_i, y_i) + \lambda \|\boldsymbol{\beta}\|_0 \right].$$

If we remove one nonzero component (backward step) from a tentative solution $\boldsymbol{\beta}$, then the term $\lambda\|\boldsymbol{\beta}\|_0$ is decreased by λ; in the mean time, if $\frac{1}{n}\sum_{i=1}^{n}\phi(\boldsymbol{\beta}^\top \mathbf{x}_i, y_i)$ is increased by an amount less than λ, then the overall regularized objective function is decreased. In this case, the backward step should be taken because it decreases the overall regularized objective value. However, a main problem of using a fixed λ is when λ is small (which is required to achieve good statistical performance), the backward step is ineffective because it can only occur after a significant number of forward steps, which have already overfitted the data. This problem is similar to that of the standard backward greedy algorithm. The FoBa algorithm, which chooses the λ threshold adaptively, fixes the problem. If we pick $\nu \to 1$, then the algorithm can be regarded as an approximate path-following scheme (similar to the LARS method for Lasso) with gradually decreasing λ. Moreover, the algorithm is also theoretically justified. Under the sparse eigenvalue condition, we obtained strong performance bounds for FoBa for feature selection and oracle inequalities. In fact, to the author's knowledge, in terms of sparsity, the bounds developed for FoBa in this paper are superior to earlier results in the literature for other methods.

Our experiments also showed that FoBa achieves its design goal: that is, it gives smaller training error than either forward-greedy or L_1 regularization for any given level of sparsity. Therefore, the experiments are consistent with our theory. In simulation, better sparsity leads to better parameter estimation accuracy. In real data, better sparsity helps on some data (e.g., *Boston Housing*) but not always. This implies that sparsity may not be the best complexity measure for any given problem. In particular, as shown in the *Ionosphere* experiment, the prior knowledge of using small coefficients, which is encoded in the L_1 regularization formulation, can lead to better generalization performance (when such a prior is appropriate for the problem). The so called "bias" of L_1 regularization, which leads to suboptimal sparsity on the training data, can be advantageous on the test data.

The experiments also indicate that in order to design learning methods with the best possible generalization performance, one should consider factors beyond sparsity. In fact, for some data, features cannot be reliably selected due to the high correlation among some key variables or the small signal to noise ratio. In this scenario, any algorithm (including FoBa) that mimics L_0-regularization is unstable (that is, it often selects incorrect features), which hurts the prediction performance. Additional prior knowledge, such as small coefficients in certain norm, can be very important. We believe such a prior knowledge can be incorporated into FoBa by adding a suitable regularization into the objective function. In particular, it should be possible to design a FoBa like path-following algorithm that simultaneously achieves sparsity and small coefficients. This important extension of FoBa will be studied in a future work.

APPENDIX A
PROOFS OF THEOREMS

Throughout the proofs, we denote by $\mathbf{v}_F$ the restriction of a vector $\mathbf{v} \in R^d$ to coefficients in a set $F \subset \{1, \ldots, d\}$.

Before going into the technical details, we shall first describe the high-level ideas behind the proofs.

First, we outline the properties of FoBa under the assumption that $X\bar{\boldsymbol{\beta}} = \mathbf{E}\mathbf{y}$ for some sparse $\bar{\boldsymbol{\beta}}$. It can be shown that the forward greedy step makes significant progress unless the objective value $Q(\boldsymbol{\beta}^{(k)})$ is not much larger than that of $Q(\bar{\boldsymbol{\beta}})$; the bound is given in Lemma B.1. Since at termination, forward step makes small progress, we can show that most features in $\mathrm{supp}(\bar{\boldsymbol{\beta}})$ have to belong to $F^{(k)}$ when FoBa terminates. Moreover, backward steps ensure that the coefficients of $\boldsymbol{\beta}^{(k)}$ at the start of any forward iteration cannot be too large; the bound is given in Lemma B.2. This fact means that if we set the threshold ϵ in the FoBa algorithm sufficiently larger than the noise level, then the nonzero coefficients of $\boldsymbol{\beta}^{(k)}$ should contain more signal than noise on average. The combination of the above two ideas means that when FoBa terminates, large nonzero coefficients of $\bar{\boldsymbol{\beta}}$ have to belong to $\mathrm{supp}(\boldsymbol{\beta}^{(k)})$, and moreover, $\boldsymbol{\beta}^{(k)}$ doesn't contain many nonzero coefficients that are purely noise (which doesn't belong to $\mathrm{supp}(\bar{\boldsymbol{\beta}})$). This observation leads to the feature selection statement in Lemma B.4 (and Theorem 3.3). A similar reasoning shows the FoBa iteration cannot reach k that is significantly larger than $\|\bar{\boldsymbol{\beta}}\|_0$, as in Theorem 3.2.

If $\mathbf{E}\mathbf{y}$ cannot be approximated by $X\bar{\boldsymbol{\beta}}$ for any sparse $\bar{\boldsymbol{\beta}}$, then we cannot prove meaningful feature selection results. However, the above mentioned bounds for forward and backward steps still imply that each increase of k will introduce more signal than noise. That is, when k is sufficiently large compared to $\bar{k}$, $Q(\boldsymbol{\beta}^{(k)})$ is significantly smaller than $Q(\bar{\boldsymbol{\beta}})$, which leads to the oracle style inequalities in Lemma B.3 (which applies during the FoBa iterations) and in Lemma B.4 (which applies when FoBa terminates). While in such case, we cannot prove feature selection results, we can still obtain a good oracle inequality when we compare $\boldsymbol{\beta}^{(k)}$ to any sparse $\bar{\boldsymbol{\beta}}$ simply because we can reduce the risk $Q(\boldsymbol{\beta}^{(k)})$ sufficiently quickly in this case, which allows us to employ a nonsparse $\boldsymbol{\beta}^{(k)}$ to compete with a sparse $\bar{\boldsymbol{\beta}}$. Since the true target is nonsparse, this comparison is to our advantage, and hence, as long as k is not too large (to avoid overfitting), it is possible to achieve an oracle inequality $\|X\boldsymbol{\beta}^{(k)} - \mathbf{E}\mathbf{y}\|_2 \leq \|X\bar{\boldsymbol{\beta}} - \mathbf{E}\mathbf{y}\|_2$ as shown in Theorem 3.1.

In order to handle the effect of noise appropriately, our analysis also requires carefully derived concentration results for sub-Gaussian noise, which are given in Appendix C. In particular, to obtain a general oracle inequality (as in Theorem 3.1), we need uniform concentration bounds for sparse least squares regression stated in Lemma C.3. In order to show that the final solution is sparse as in Theorem 3.2 and Theorem 3.3 (under the assumption that $\|X\hat{\boldsymbol{\beta}}(\bar{F}, \mathbf{E}\mathbf{y}) - \mathbf{E}\mathbf{y}\|_2$ is small), we need more subtle noise concentration estimates in Lemma C.4 and Lemma C.5.

A. Proof of Theorem 3.1

For convenience, we first state the following simple algebraic result.

Proposition A.1: Given $a, b, c > 0$. If $a^2 \geq 20c$, then $-a^2/2 + 2a\sqrt{b+c} \leq 10b$.

Proof: If $b \geq c/4$, then $-a^2/2 + 2a\sqrt{b+c} \leq -a^2/2 + 2a\sqrt{5b} \leq 10b$, where the second inequality is obtained by taking maximum over a achieved at $a = 2\sqrt{5b}$. Otherwise, we have $b < c/4$, and thus, $-a^2/2 + 2a\sqrt{b+c} \leq -a^2/2 + 2a\sqrt{(5/4)c} \leq -a^2/2 + 2a\sqrt{(5/4)a^2/20} = 0$, where the assumption $a^2 \geq 20c$ is used. ∎

Now we are ready to prove the theorem. With probability $1 - \eta$, the event in Lemma C.3 holds. That is, for all k

$$\begin{aligned} &\left\|X\boldsymbol{\beta}^{(k)} - \mathbf{E}\mathbf{y}\right\|_2^2 - \|X\bar{\boldsymbol{\beta}} - \mathbf{E}\mathbf{y}\|_2^2 \\ &\leq n\left[Q\left(\boldsymbol{\beta}^{(k)}\right) - Q(\bar{\boldsymbol{\beta}})\right] + 2\left\|X\boldsymbol{\beta}^{(k)} - X\bar{\boldsymbol{\beta}}\right\|_2 \\ &\quad \cdot \sqrt{7.4|\bar{F}| + 2.7|F^{(k)} - \bar{F}|\ln(16d) + 2.7\ln(2e/\eta)}\sigma. \end{aligned} \tag{7}$$

We first consider the situation that $k \leq s - \bar{k}$. In this situation, we may consider the two cases indicated in Lemma B.4. In the first case, the second inequality of Lemma B.4 holds, and in the second case, the third displayed inequality of Lemma B.4 holds. In both cases, we have the following inequality [see (8), shown at the bottom of the page]. Now by adding this inequality to (7), we obtain

$$\begin{aligned} &\left\|X\boldsymbol{\beta}^{(k)} - \mathbf{E}\mathbf{y}\right\|_2^2 - \|X\bar{\boldsymbol{\beta}} - \mathbf{E}\mathbf{y}\|_2^2 - 18n\epsilon\rho(s)^{-1}\Delta k(\epsilon, s) \\ &\leq -\left\|X\bar{\boldsymbol{\beta}} - X\boldsymbol{\beta}^{(k)}\right\|_2^2 \Big/ 2 + 2\left\|X\boldsymbol{\beta}^{(k)} - X\bar{\boldsymbol{\beta}}\right\|_2 \\ &\quad \cdot \sqrt{7.4|\bar{F}| + 2.7|F^{(k)} - \bar{F}|\ln(16d) + 2.7\ln(2e/\eta)}\sigma. \end{aligned}$$

Let $a = \|X\bar{\boldsymbol{\beta}} - X\boldsymbol{\beta}^{(k)}\|_2$, $b = 7.4\sigma^2|\bar{F}| + 2.7\sigma^2\ln(2e/\eta)$, and $c = 2.7|F^{(k)} - \bar{F}|\sigma^2\ln(16d)$. From the first inequality of Lemma B.4, we obtain $a^2 \geq n\rho(s)|F^{(k)} - \bar{F}|\epsilon/2 \geq 54|F^{(k)} - \bar{F}|\sigma^2\ln(16d) = 20c$, where the condition on ϵ in the theorem statement is used. Therefore, we obtain from Proposition A.1

$$\begin{aligned} &\left\|X\boldsymbol{\beta}^{(k)} - \mathbf{E}\mathbf{y}\right\|_2^2 - \|X\bar{\boldsymbol{\beta}} - \mathbf{E}\mathbf{y}\|_2^2 \\ &\qquad \leq 74\sigma^2|\bar{F}| + 27\sigma^2\ln(2e/\eta) + 18n\epsilon\rho(s)^{-1}\Delta k(\epsilon, s). \end{aligned}$$

That is, the first displayed inequality of the theorem holds.

Next, we consider the situation that when FoBa terminates, $k \geq s - \bar{k}$. For each $k_0 \in [0.8s - \bar{k}, s - \bar{k}]$, when FoBa first reaches $k = k_0$, the immediate previous iteration does not have

$$n\left[Q\left(\boldsymbol{\beta}^{(k)}\right) - Q(\bar{\boldsymbol{\beta}})\right] + \left\|X\bar{\boldsymbol{\beta}} - X\boldsymbol{\beta}^{(k)}\right\|_2^2 \Big/ 2 \leq 18n\epsilon\rho(s)^{-1}\Delta k(\epsilon, s) \tag{8}$$

backward step, and the assumption of the theorem on s implies that

$$2\gamma^2\left|\left(\bar{F}\cup F^{(k)}\right)-F^{(k-1)}\right| \le 32(\bar{k}+1)\le(0.8s-2\bar{k})\rho(s)^2\le\left|F^{(k)}-\bar{F}\right|\rho(s)^2 \quad (9)$$

with $\gamma=4$. Therefore, Lemma B.3 applies, and it implies that (note that $s\ge|\bar{F}\cup F^{(k)}|$)

$$n\left[Q\left(\boldsymbol{\beta}^{(k)}\right)-Q(\bar{\boldsymbol{\beta}})\right]\le -0.5\left\|X\bar{\boldsymbol{\beta}}-X\boldsymbol{\beta}^{(k)}\right\|_2^2 \quad (10)$$

$$n^{-1}\left\|X\left(\bar{\boldsymbol{\beta}}-\boldsymbol{\beta}^{(k)}\right)\right\|_2^2\ge\rho(s)\left|F^{(k)}-\bar{F}\right|\epsilon\Big/2 \ge 54\left|F^{(k)}-\bar{F}\right|\sigma^2\ln(16d)\Big/n. \quad (11)$$

The last inequality uses the condition of ϵ in the theorem. We thus have (note that $\rho(s)\le 1$ for all s since the basis functions are all normalized in Assumption 3.1)

$$\begin{aligned}
&\left\|X\boldsymbol{\beta}^{(k)}-\mathbf{E}\mathbf{y}\right\|_2^2-\|X\bar{\boldsymbol{\beta}}-\mathbf{E}\mathbf{y}\|_2^2\\
&\le-\left\|X\bar{\boldsymbol{\beta}}-X\boldsymbol{\beta}^{(k)}\right\|_2^2\Big/2+2\left\|X\boldsymbol{\beta}^{(k)}-X\bar{\boldsymbol{\beta}}\right\|_2\\
&\quad\cdot\sqrt{7.4|\bar{F}|+2.7|F^{(k)}-\bar{F}|\ln(16d)+2.7\ln(2e/\eta)}\sigma\\
&\le-\left\|X\bar{\boldsymbol{\beta}}-X\boldsymbol{\beta}^{(k)}\right\|_2^2\Big/2+2\left\|X\boldsymbol{\beta}^{(k)}-X\bar{\boldsymbol{\beta}}\right\|_2\\
&\quad\cdot\sqrt{(2.7/32+2.7)|F^{(k)}-\bar{F}|\ln(16d)+2.7\ln(2e/\eta)}\sigma\\
&\le-\left\|X\bar{\boldsymbol{\beta}}-X\boldsymbol{\beta}^{(k)}\right\|_2^2\Big/2+2\left\|X\boldsymbol{\beta}^{(k)}-X\bar{\boldsymbol{\beta}}\right\|_2\\
&\quad\cdot\sqrt{(1.65/32)\left\|X(\bar{\boldsymbol{\beta}}-\boldsymbol{\beta}^{(k)})\right\|_2^2+2.7\ln(2e/\eta)\sigma^2}\\
&\le-\left\|X\bar{\boldsymbol{\beta}}-X\boldsymbol{\beta}^{(k)}\right\|_2^2\Big/2+\left\|X\boldsymbol{\beta}^{(k)}-X\bar{\boldsymbol{\beta}}\right\|_2^2\Big/4\\
&\quad+4\left[(1.65/32)\left\|X(\bar{\boldsymbol{\beta}}-\boldsymbol{\beta}^{(k)})\right\|_2^2+2.7\ln(2e/\eta)\sigma^2\right]\\
&\le-(0.35/8)\left\|X\bar{\boldsymbol{\beta}}-X\boldsymbol{\beta}^{(k)}\right\|_2^2+10.8\ln(2e/\eta)\sigma^2\\
&\le-(0.35/16)n\rho(s)\left|F^{(k)}-\bar{F}\right|\epsilon+10.8\ln(2e/\eta)\sigma^2\\
&\le-0.7n\rho(s)^{-1}|\bar{F}|\epsilon+10.8\ln(2e/\eta)\sigma^2
\end{aligned}$$

where the first inequality is obtained by adding (10) to (7). The second inequality uses $7.4|\bar{F}|=7.4\bar{k}\le 2.7\bar{k}\ln(16d)\le 2.7\rho(s)^2|F^{(k)}-\bar{F}|\ln(16d)/32$ according to (9), and $\rho(s)\le 1$. The third inequality uses (11). The fourth inequality uses $2ab\le a^2/4+4b^2$. The fifth inequality is simple algebra. The sixth inequality uses (11). The last inequality uses (9), which implies that $\rho(s)|F^{(k)}-\bar{F}|\ge 32\rho(s)^{-1}|\bar{F}|$. This implies the theorem by noticing that $\rho(s)\le 1$.

B. Proof of Theorem 3.2

With probability $1-2\eta$, we assume that both probability events in Lemma C.3 and in Lemma C.4 hold.

First we show that at any time during the FoBa algorithm, we must have $k<s-\bar{k}$. Assume this is not true, then consider the first time we reach $k+\bar{k}=s$. In this case, we know that there is no backward step in the immediate previous iteration, and with $\gamma=2$, we have from the assumption of the theorem on s

$$\begin{aligned}
&2\gamma^2\left|\left(\bar{F}\cup F^{(k)}\right)-F^{(k-1)}\right|\\
&\le 8(\bar{k}+1)\le(s-2\bar{k})\rho(s)^2\le\left|F^{(k)}-\bar{F}\right|\rho(s)^2.
\end{aligned}$$

Therefore, Lemma B.3 applies with $\gamma=2$. It means that (here we take $\bar{\boldsymbol{\beta}}$ in Lemma B.3 as $\hat{\boldsymbol{\beta}}(\bar{F})$)

$$\begin{aligned}
&Q\left(\hat{\boldsymbol{\beta}}\left(F^{(k)}\cup\bar{F}\right)\right)\\
&\quad\le Q(\hat{\boldsymbol{\beta}}(\bar{F}))-(1-0.5)^2\rho(s)\left|F^{(k)}-\bar{F}\right|\epsilon/2. \quad (12)
\end{aligned}$$

Moreover, Lemma C.4 implies that

$$\begin{aligned}
\sqrt{n\left[Q(\hat{\boldsymbol{\beta}}(\bar{F}))-Q\left(\hat{\boldsymbol{\beta}}\left(F^{(k)}\cup\bar{F}\right)\right)\right]}&\le\|X\hat{\boldsymbol{\beta}}(\bar{F},\mathbf{E}\mathbf{y})-\mathbf{E}\mathbf{y}\|_2\\
&\quad+\sigma\sqrt{2.7|F^{(k)}-\bar{F}|\ln(16d)+2.7\ln(2e/\eta)}.
\end{aligned}$$

It implies that

$$\begin{aligned}
n\left[Q(\hat{\boldsymbol{\beta}}(\bar{F}))-Q\left(\hat{\boldsymbol{\beta}}(F^{(k)}\cup\bar{F})\right)\right]&\le 2\|X\hat{\boldsymbol{\beta}}(\bar{F},\mathbf{E}\mathbf{y})-\mathbf{E}\mathbf{y}\|_2^2\\
&\quad+5.4\sigma^2\left[\left|F^{(k)}-\bar{F}\right|\ln(16d)+\ln(2e/\eta)\right]. \quad (13)
\end{aligned}$$

By adding (13) to n times (12), we obtain

$$\begin{aligned}
n\rho(s)\left|F^{(k)}-\bar{F}\right|\epsilon\Big/8&\le 2\left\|X\hat{\boldsymbol{\beta}}(\bar{F},\mathbf{E}\mathbf{y})-\mathbf{E}\mathbf{y}\right\|_2^2\\
&\quad+5.4\sigma^2\left[\left|F^{(k)}-\bar{F}\right|\ln(16d)+\ln(2e/\eta)\right].
\end{aligned}$$

This inequality can be rewritten as

$$\begin{aligned}
n\epsilon&\le\frac{8\rho(s)^{-1}}{|F^{(k)}-\bar{F}|}\left[2\|X\hat{\boldsymbol{\beta}}(\bar{F},\mathbf{E}\mathbf{y})-\mathbf{E}\mathbf{y}\|_2^2\right.\\
&\quad\left.+5.4\sigma^2\left[\left|F^{(k)}-\bar{F}\right|\ln(16d)+\ln(2e/\eta)\right]\right]\\
&\le\left[\frac{2\rho(s)}{\bar{k}+1}\|X\hat{\boldsymbol{\beta}}(\bar{F},\mathbf{E}\mathbf{y})-\mathbf{E}\mathbf{y}\|_2^2\right.\\
&\quad\left.+5.4\sigma^2[8\rho(s)^{-1}\ln(16d)+\rho(s)(\bar{k}+1)^{-1}\ln(2e/\eta)]\right].
\end{aligned}$$

The second inequality uses the assumption $\bar{k}+1\le(s-2\bar{k})\rho(s)^2/8\le|F^{(k)}-\bar{F}|\rho(s)^2/8$. However, the above displayed inequality is a contradiction to the assumption of ϵ. Therefore, the FoBa algorithm cannot reach $k=s-\bar{k}$.

The proof of the oracle inequality, under the condition $\epsilon\ge 108\rho(s)^{-1}\sigma^2\ln(16d)/n$, is identical to the derivation of the same oracle inequality in Theorem 3.1.

C. Proof of Theorem 3.3

With probability $1-3\eta$, all probability events in Lemma C.5, Lemma C.3, and Lemma C.4 hold. The proof of Theorem 3.2 implies that the FoBa algorithm terminates at $k<s-\bar{k}$.

At the end of FoBa algorithm, we can apply Lemma B.4, which implies two situations, where we take $\bar{\boldsymbol{\beta}}$ in the statement of the lemma as $\hat{\boldsymbol{\beta}}(\bar{F})$.

The first situation is when

$$\begin{aligned} nQ\left(\hat{\boldsymbol{\beta}}\left(\bar{F}\cup F^{(k)}\right)\right) &\leq nQ\left(\boldsymbol{\beta}^{(k)}\right) \\ &\leq nQ(\hat{\boldsymbol{\beta}}(\bar{F})) - \left\|X\hat{\boldsymbol{\beta}}(\bar{F}) - X\boldsymbol{\beta}^{(k)}\right\|_2^2 \Big/ 2 \\ &\leq nQ(\hat{\boldsymbol{\beta}}(\bar{F})) - n\rho(s)\left|F^{(k)} - \bar{F}\right|\epsilon \Big/ 2. \end{aligned} \tag{14}$$

Note that $X\hat{\boldsymbol{\beta}}(\bar{F}, \mathbf{E}\mathbf{y}) = \mathbf{E}\mathbf{y}$ by the assumption of the theorem. We thus obtain from Lemma C.4

$$\begin{aligned} n\left[Q(\hat{\boldsymbol{\beta}}(\bar{F})) - Q\left(\hat{\boldsymbol{\beta}}\left(F^{(k)}\cup\bar{F}\right)\right)\right] \\ \leq \sigma^2\left[2.7\left|F^{(k)} - \bar{F}\right|\ln(16d) + 2.7\ln(2e/\eta)\right]. \end{aligned} \tag{15}$$

Adding this inequality to (14), we obtain

$$\begin{aligned} 0 \leq -n\rho(s)\left|F^{(k)} - \bar{F}\right|\epsilon \Big/ 2 \\ +2.7\sigma^2\left[\left|F^{(k)} - \bar{F}\right|\ln(16d) + \ln(2e/\eta)\right]. \end{aligned}$$

Using the assumption of ϵ in the theorem, we obtain (note that $\rho(s) \leq 1$ for all s since the basis functions are all normalized in Assumption 3.1)

$$\begin{aligned} 2.7\left|F^{(k)} - \bar{F}\right|\sigma^2[8\ln(16d) + 2\ln(2e/\eta)] \\ < 2.7\sigma^2\left[\left|F^{(k)} - \bar{F}\right|\ln(16d) + \ln(2e/\eta)\right]. \end{aligned}$$

This inequality can only be satisfied when

$$\left|F^{(k)} - \bar{F}\right| = 0.$$

Moreover, we also have

$$\begin{aligned} &n^{-1}\left\|X\hat{\boldsymbol{\beta}}(\bar{F}) - X\boldsymbol{\beta}^{(k)}\right\|_2^2 \\ &\quad\geq \rho(s)\left\|\hat{\boldsymbol{\beta}}(\bar{F}) - \boldsymbol{\beta}^{(k)}\right\|_2^2 \geq \rho(s)\left\|\hat{\boldsymbol{\beta}}(\bar{F})_{\bar{F}-F^{(k)}}\right\|_2^2 \\ &\quad\geq \rho(s)\left[0.5\left\|\bar{\boldsymbol{\beta}}_{\bar{F}-F^{(k)}}\right\|_2^2 - \left\|(\bar{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(\bar{F}))_{\bar{F}-F^{(k)}}\right\|_2^2\right] \\ &\quad\geq 0.5\rho(s)\left\|\bar{\boldsymbol{\beta}}_{\bar{F}-F^{(k)}}\right\|_2^2 - \rho(s)\left|\bar{F} - F^{(k)}\right|\|\bar{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(\bar{F})\|_\infty^2 \end{aligned} \tag{16}$$

where the first inequality uses the definition of $\rho(s)$ in Definition 3.1. The second inequality uses the fact that $\beta_j^{(k)} = 0$ when $j \in \bar{F} - F^{(k)}$. The third inequality uses the algebraic inequality $\|a\|_2^2 \geq 0.5\|b\|_2^2 - \|a-b\|_2^2$. The fourth inequality uses the upper bound of a vector's 2-norm by its ∞-norm.

Since we have already proved that $|F^{(k)} - \bar{F}| = 0$, we can also add (14) to (15) and obtain

$$\begin{aligned} &n^{-1}\left\|X\hat{\boldsymbol{\beta}}(\bar{F}) - X\boldsymbol{\beta}^{(k)}\right\|_2^2 \\ &\quad\leq (2n^{-1})2.7\sigma^2\left[\left|F^{(k)} - \bar{F}\right|\ln(16d) + \ln(2e/\eta)\right] \\ &\quad= 5.4\sigma^2\ln(2e/\eta)/n. \end{aligned}$$

If $|\bar{F} - F^{(k)}| = 0$, then the claim of the theorem holds automatically since we have already shown $|F^{(k)} - \bar{F}| = 0$. Otherwise, by subtracting (16) from the above inequality, we obtain

$$\begin{aligned} &0.5\rho(s)\left\|\bar{\boldsymbol{\beta}}_{\bar{F}-F^{(k)}}\right\|_2^2 \\ &\leq \rho(s)\left|\bar{F} - F^{(k)}\right|\|\bar{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(\bar{F})\|_\infty^2 + 5.4\sigma^2\ln(2e/\eta)\Big/ n \\ &\leq 2\rho(s)\left|\bar{F} - F^{(k)}\right|\sigma^2\ln(2\bar{k}/\eta))/(n\rho(\bar{k})) + 5.4\sigma^2\ln(2e/\eta)\Big/ n \\ &\leq 5.4|\bar{F} - F^{(k)}|\epsilon. \end{aligned}$$

In the above derivation, the second inequality uses Lemma C.5 and $\bar{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\bar{F}, \mathbf{E}\mathbf{y})$ (which follows from $\mathbf{E}\mathbf{y} = X\bar{\boldsymbol{\beta}}$). The third inequality uses $|\bar{F} - F^{(k)}| \geq 1$, $\rho(s) \leq 1$, and the assumption of ϵ in the theorem. This inequality implies that

$$\begin{aligned} 0.5\rho(s)\left|\left\{j \in \bar{F} - F^{(k)} : \bar{\boldsymbol{\beta}}_j^2 \geq 32\epsilon\rho(s)^{-2}\right\}\right| 32\epsilon\rho(s)^{-2} \\ \leq 0.5\rho(s)\left\|\bar{\boldsymbol{\beta}}_{\bar{F}-F^{(k)}}\right\|_2^2 \leq 5.4\left|\bar{F} - F^{(k)}\right|\epsilon. \end{aligned}$$

Therefore, we have

$$\left|\left\{j \in \bar{F} - F^{(k)} : \bar{\boldsymbol{\beta}}_j^2 \geq 32\epsilon\rho(s)^{-2}\right\}\right| \leq |\bar{F} - F^{(k)}|/2$$

and thus

$$\begin{aligned} \left|\bar{F} - F^{(k)}\right| &\leq 2|\{j \in \bar{F} - F^{(k)} : \bar{\boldsymbol{\beta}}_j^2 < 32\epsilon\rho(s)^{-2}\}| \\ &\leq 2\left|\left\{j \in \bar{F} : \bar{\boldsymbol{\beta}}_j^2 < 32\epsilon\rho(s)^{-2}\right\}\right|. \end{aligned}$$

This shows that in the first situation (where (14) holds), the conclusion of the theorem is valid.

In the second situation, we assume that (14) does not hold. In this case, we may simply apply the second conclusion of Lemma B.4 that implies

$$\begin{aligned} \rho(s)\left|F^{(k)} - \bar{F}\right|\epsilon/2 &\leq n^{-1}\left\|X\left(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(k)}\right)\right\|_2^2 \\ &\leq 16\epsilon\rho(s)^{-1}\left|\bar{F} - F^{(k)}\right| \\ &\leq 32\epsilon\rho(s)^{-1}\left|\left\{j \in \bar{F} : \bar{\boldsymbol{\beta}}_j^2 < 32\epsilon\rho(s)^{-2}\right\}\right|. \end{aligned}$$

This automatically implies the desired bound.

Appendix B
Properties of FoBa

For convenience, we state a simple fact of least squares solution.

Proposition B.1: Let $F \subset \{1, \ldots, d\}$, $\mathbf{g} \in R^n$, and let $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}(F, \mathbf{g})$. Then for all $\boldsymbol{\beta}' \in R^d$

$$\begin{aligned} &\|X\boldsymbol{\beta}' - \mathbf{g}\|_2^2 - \|X\boldsymbol{\beta} - \mathbf{g}\|_2^2 \\ &\quad= \|X(\boldsymbol{\beta}' - \boldsymbol{\beta})\|_2^2 + 2(X\boldsymbol{\beta} - \mathbf{g})^\top \sum_{j\in\mathrm{supp}(\boldsymbol{\beta}')-F} (\boldsymbol{\beta}'_j - \boldsymbol{\beta}_j)\mathbf{f}_j. \end{aligned}$$

In particular, if $\mathrm{supp}(\boldsymbol{\beta}') \subset F$, then

$$\|X\boldsymbol{\beta}' - \mathbf{g}\|_2^2 - \|X\boldsymbol{\beta} - \mathbf{g}\|_2^2 = \|X(\boldsymbol{\beta}' - \boldsymbol{\beta})\|_2^2.$$

Proof: The optimality of $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}(F, \mathbf{g})$ as the least squares solution in F implies that for all $j \in F$

$$\mathbf{f}_j^\top (X\boldsymbol{\beta} - \mathbf{g}) = 0.$$

Therefore, if we let $F' = \mathrm{supp}(\boldsymbol{\beta}')$, then

$$\begin{aligned}
&2(X\boldsymbol{\beta} - \mathbf{g})^\top \sum_{j \in F' - F} (\boldsymbol{\beta}'_j - \boldsymbol{\beta}_j)\mathbf{f}_j \\
&\quad = 2(X\boldsymbol{\beta} - \mathbf{g})^\top \sum_{j \in F' \cup F} (\boldsymbol{\beta}'_j - \boldsymbol{\beta}_j)\mathbf{f}_j \\
&\quad = 2(X\boldsymbol{\beta} - \mathbf{g})^\top (X\boldsymbol{\beta}' - X\boldsymbol{\beta}) \\
&\quad = -\|X\boldsymbol{\beta}' - X\boldsymbol{\beta}\|_2^2 + \|X\boldsymbol{\beta}' - \mathbf{g}\|_2^2 - \|X\boldsymbol{\beta} - \mathbf{g}\|_2^2.
\end{aligned}$$

This implies the proposition. ∎

The following lemma provides a lower bound on the squared error reduction of one forward greedy step.

Lemma B.1: Let Assumption 3.1 hold. Consider any $\boldsymbol{\beta}' \in R^d$. Consider $F' = \mathrm{supp}(\boldsymbol{\beta}')$ and $F \subset \{1, \ldots, d\}$. Let $s = |F' \cup F|$. Let $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}(F)$. If for some $\alpha \geq -1$, we have

$$Q(\boldsymbol{\beta}) - Q(\boldsymbol{\beta}') = \frac{\alpha}{n}\|X(\boldsymbol{\beta} - \boldsymbol{\beta}')\|_2^2$$

then

$$\begin{aligned}
\inf_{\alpha \in R, j \in F' - F} Q(\boldsymbol{\beta} + \alpha \mathbf{e}_j) \leq Q(\boldsymbol{\beta}) \\
- \frac{\rho(s)(1+\alpha)}{4|F' - F|}\left(\frac{1}{n}\|X(\boldsymbol{\beta} - \boldsymbol{\beta}')\|_2^2 + Q(\boldsymbol{\beta}) - Q(\boldsymbol{\beta}')\right).
\end{aligned}$$

Proof: We have from Proposition B.1 (with $\mathbf{g} = \mathbf{y}$)

$$\begin{aligned}
&2(X\boldsymbol{\beta} - \mathbf{y})^\top \sum_{j \in F' - F} (\boldsymbol{\beta}'_j - \boldsymbol{\beta}_j)\mathbf{f}_j \\
&\qquad = -\|X(\boldsymbol{\beta}' - \boldsymbol{\beta})\|_2^2 + n[Q(\boldsymbol{\beta}') - Q(\boldsymbol{\beta})]. \quad (17)
\end{aligned}$$

This leads to the following derivation for an arbitrary fixed $\eta > 0$ (which we will pick to optimize the bound later on)

$$\begin{aligned}
&|F' - F| \inf_{j \in F' - F} Q(\boldsymbol{\beta} + \eta(\boldsymbol{\beta}'_j - \boldsymbol{\beta}_j)\mathbf{e}_j) \\
&\leq \sum_{j \in F' - F} Q(\boldsymbol{\beta} + \eta(\boldsymbol{\beta}'_j - \boldsymbol{\beta}_j)\mathbf{e}_j) \\
&= |F' - F|Q(\boldsymbol{\beta}) + \frac{\eta^2}{n}\sum_{j \in F' - F} (\boldsymbol{\beta}'_j - \boldsymbol{\beta}_j)^2 \|\mathbf{f}_j\|_2^2 \\
&\qquad + \frac{2\eta}{n}(X\boldsymbol{\beta} - \mathbf{y})^\top \sum_{j \in F' - F} (\boldsymbol{\beta}'_j - \boldsymbol{\beta}_j)\mathbf{f}_j \\
&= |F' - F|Q(\boldsymbol{\beta}) + \eta^2 \sum_{j \in F' - F} (\boldsymbol{\beta}'_j - \boldsymbol{\beta}_j)^2 \\
&\qquad - \eta\left[\frac{1}{n}\|X(\boldsymbol{\beta}' - \boldsymbol{\beta})\|_2^2 + Q(\boldsymbol{\beta}) - Q(\boldsymbol{\beta}')\right]. \quad (18)
\end{aligned}$$

In the above derivation, the first inequality is simple algebra. The first equality uses the definition of $Q(\cdot)$ as squared loss and simple algebra. The second equality uses $\|\mathbf{f}_j\|_2^2 = n$ in Assumption 3.1, as well as (17). Let $\Delta Q = \frac{1}{n}\|X(\boldsymbol{\beta}' - \boldsymbol{\beta})\|_2^2 + Q(\boldsymbol{\beta}) - Q(\boldsymbol{\beta}')$. Then by optimizing over η, we obtain

$$\begin{aligned}
&|F' - F| \inf_{\eta} \inf_{j \in F' - F} Q(\boldsymbol{\beta} + \eta(\boldsymbol{\beta}'_j - \boldsymbol{\beta}_j)\mathbf{e}_j) \\
&\leq |F' - F|Q(\boldsymbol{\beta}) - \frac{[\frac{1}{n}\|X(\boldsymbol{\beta}' - \boldsymbol{\beta})\|_2^2 + Q(\beta) - Q(\boldsymbol{\beta}')]^2}{4\sum_{j \in F'}(\boldsymbol{\beta}'_j - \boldsymbol{\beta}_j)^2} \\
&= |F' - F|Q(\boldsymbol{\beta}) - \frac{(1+\alpha)n^{-1}\|X(\boldsymbol{\beta}' - \boldsymbol{\beta})\|_2^2}{4\sum_{j \in F'}(\boldsymbol{\beta}'_j - \boldsymbol{\beta}_j)^2}\Delta Q \\
&\leq |F' - F|Q(\boldsymbol{\beta}) - \frac{(1+\alpha)\rho(s)\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2}{4\sum_{j \in F'}(\boldsymbol{\beta}'_j - \boldsymbol{\beta}_j)^2}\Delta Q \\
&\leq |F' - F|Q(\boldsymbol{\beta}) - \frac{\rho(s)(1+\alpha)}{4}\Delta Q.
\end{aligned}$$

The first inequality is obtained by optimizing η from (18). The first equality uses the definition of α in the lemma. The second inequality uses the definition of $\rho(s)$ in Definition 3.1. The last inequality uses the fact that $\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2 \geq \sum_{j \in F'}(\boldsymbol{\beta}'_j - \boldsymbol{\beta}_j)^2$. This leads to the lemma. ∎

The following lemma provides an upper bound on the squared error increase of one backward greedy step.

Lemma B.2: Let Assumption 3.1 hold. Consider $\boldsymbol{\beta}' \in R^d$ and let $\bar{F} = \mathrm{supp}(\boldsymbol{\beta}')$. Consider $F \subset \{1, \ldots, d\}$ and let $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}(F)$. Let $s = |F \cup \bar{F}|$. Then

$$\inf_{j \in F} Q(\boldsymbol{\beta} - \boldsymbol{\beta}_j \mathbf{e}_j) \leq Q(\boldsymbol{\beta}) + \frac{1}{|F - \bar{F}|}\sum_{j \in F - \bar{F}} \boldsymbol{\beta}_j^2.$$

Proof: For all $j \in F$, we have $Q(\boldsymbol{\beta} + \alpha \mathbf{e}_j)$ achieves minimum at $\alpha = 0$. This implies that $(X\boldsymbol{\beta} - \mathbf{y})^\top \mathbf{f}_j = 0$ for $j \in F$. We thus have

$$\begin{aligned}
&|F - \bar{F}| \inf_{j \in F} nQ(\boldsymbol{\beta} - \boldsymbol{\beta}_j \mathbf{e}_j) \\
&\leq \sum_{j \in F - \bar{F}} nQ(\boldsymbol{\beta} - \boldsymbol{\beta}_j \mathbf{e}_j) \\
&= \sum_{j \in F - \bar{F}} \|X\boldsymbol{\beta} - \mathbf{y} - \boldsymbol{\beta}_j \mathbf{f}_j\|_2^2 \\
&= |F - \bar{F}|\|X\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \sum_{j \in F - \bar{F}} \boldsymbol{\beta}_j^2 \|\mathbf{f}_j\|_2^2 \\
&= |F - \bar{F}|nQ(\boldsymbol{\beta}) + n\sum_{j \in F - \bar{F}} \boldsymbol{\beta}_j^2.
\end{aligned}$$

The first inequality and the first equality use simple algebra. The second equality uses simple algebra and the fact that $(X\boldsymbol{\beta} - \mathbf{y})^\top \mathbf{f}_j = 0$ for $j \in F - \bar{F}$. The third equality uses $\|\mathbf{f}_j\|_2^2 = n$ in Assumption 3.1. This leads to the lemma. ∎

The following bound means that during the FoBa iterations, if the vector $\boldsymbol{\beta}^{(k)}$ becomes significantly less sparse than a competing target $\bar{\boldsymbol{\beta}}$, then $Q(\boldsymbol{\beta}^{(k)})$ is significantly smaller than $Q(\bar{\boldsymbol{\beta}})$. This fact implies an oracle inequality (see Theorem 3.1). Moreover, if $\mathbf{E}\mathbf{y}$ can be approximated by a sparse target, which means that $Q(\boldsymbol{\beta}^{(k)})$ cannot be too smaller than $Q(\bar{\boldsymbol{\beta}})$, the result implies

that FoBa will terminate at a point k that is not much larger than $\bar{k}$ (see Theorem 3.2).

Lemma B.3: Consider any $\bar{\boldsymbol{\beta}} \in R^d$ and let $\bar{F} = \mathrm{supp}(\bar{\boldsymbol{\beta}})$. In Fig. 4, assume that at the start of an iteration's forward step, the immediate previous iteration had taken no backward steps. If with $s = |F^{(k)} \cup \bar{F}|$ the following inequality holds for some $\gamma \geq 2$

$$2\gamma^2 \left| \left(\bar{F} \cup F^{(k)} \right) - F^{(k-1)} \right| \leq \left| F^{(k)} - \bar{F} \right| \rho(s)^2$$

then

$$n^{-1} \left\| X \left(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(k)} \right) \right\|_2^2 \geq \rho(s) \left| F^{(k)} - \bar{F} \right| \epsilon \Big/ 2$$

and

$$Q\left(\boldsymbol{\beta}^{(k)}\right) \leq Q(\bar{\boldsymbol{\beta}}) - (1 - 2\gamma^{-1}) n^{-1} \left\| X\bar{\boldsymbol{\beta}} - X\boldsymbol{\beta}^{(k)} \right\|_2^2$$

and

$$Q\left(\hat{\boldsymbol{\beta}}(F^{(k)} \cup \bar{F})\right) \leq Q(\bar{\boldsymbol{\beta}}) - (1-\gamma^{-1})^2 \rho(s) \left| F^{(k)} - \bar{F} \right| \epsilon \Big/ 2.$$

Proof: Let $F' = \bar{F} \cup F^{(k)}$. Since no backward step was taken in the immediate previous iteration, we must have (from the previous forward step) $\mathrm{supp}(\boldsymbol{\beta}^{(k-1)}) \subset F^{(k-1)} \subset \{i^{(k-1)}\} \cup F^{(k-1)} = F^{(k)} \subset F'$ by the design of the algorithm.

If we let $\boldsymbol{\beta}' = \hat{\boldsymbol{\beta}}(F')$, then Proposition B.1 implies that

$$Q\left(\boldsymbol{\beta}^{(k-1)}\right) - Q(\boldsymbol{\beta}') = n^{-1} \left\| X \left(\boldsymbol{\beta}' - \boldsymbol{\beta}^{(k-1)} \right) \right\|_2^2$$
$$Q\left(\boldsymbol{\beta}^{(k)}\right) - Q(\boldsymbol{\beta}') = n^{-1} \left\| X \left(\boldsymbol{\beta}' - \boldsymbol{\beta}^{(k-1)} \right) \right\|_2^2.$$

By definition, we have $\min_i \min_\alpha Q(\boldsymbol{\beta}^{(k-1)} + \alpha \mathbf{e}_i) = Q(\boldsymbol{\beta}^{(k-1)}) - \delta^{(k)}$. Therefore, by Lemma B.1 (apply the lemma with $\alpha = 1$, $\boldsymbol{\beta} = \boldsymbol{\beta}^{(k-1)}$, and $F = F^{(k-1)}$), we obtain for the immediate previous forward iteration

$$\begin{aligned} \delta^{(k)} &\geq \frac{\rho(s)}{\left|F' - F^{(k-1)}\right|} \left[Q\left(\boldsymbol{\beta}^{(k-1)}\right) - Q(\boldsymbol{\beta}') \right] \\ &\geq \frac{\rho(s)}{\left|F' - F^{(k-1)}\right|} \left[Q\left(\boldsymbol{\beta}^{(k)}\right) - Q(\boldsymbol{\beta}') \right] \\ &= \frac{\rho(s)}{n\left|F' - F^{(k-1)}\right|} \left\| X\boldsymbol{\beta}^{(k)} - X\boldsymbol{\beta}' \right\|_2^2 \end{aligned} \tag{19}$$

where we have used $Q(\boldsymbol{\beta}^{(k)}) \leq Q(\boldsymbol{\beta}^{(k-1)})$ in the second inequality, and $Q(\boldsymbol{\beta}^{(k)}) - Q(\boldsymbol{\beta}') = n^{-1}\|X\boldsymbol{\beta}' - X\boldsymbol{\beta}^{(k)}\|_2^2$ in the last equality.

Moreover, the backward step termination condition requires $Q(\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}_j^{(k)}\mathbf{e}_j) - Q(\boldsymbol{\beta}^{(k)}) > 0.5\delta^{(k)}$. We thus have from Lemma B.2 (apply the lemma with $F = F^{(k)}$, $\boldsymbol{\beta} = \boldsymbol{\beta}^{(k)}$, and $\boldsymbol{\beta}' = \bar{\boldsymbol{\beta}}$):

$$\begin{aligned} &\frac{1}{|F^{(k)} - \bar{F}|} \left\| \boldsymbol{\beta}^{(k)}_{F^{(k)} - \bar{F}} \right\|_2^2 \\ &= \frac{1}{|F^{(k)} - \bar{F}|} \sum_{j \in F^{(k)} - \bar{F}} \left(\boldsymbol{\beta}_j^{(k)} \right)^2 \geq \delta^{(k)}/2. \end{aligned} \tag{20}$$

Now we can derive the first desired bound of the lemma as follows:

$$\begin{aligned} n^{-1} \left\| X\bar{\boldsymbol{\beta}} - X\boldsymbol{\beta}^{(k)} \right\|_2^2 &\geq \rho(s) \left\| \bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(k)} \right\|_2^2 \\ &\geq \rho(s) \left\| \boldsymbol{\beta}^{(k)}_{F^{(k)} - \bar{F}} \right\|_2^2 \\ &\geq \rho(s) \left| F^{(k)} - \bar{F} \right| \epsilon/2 \end{aligned}$$

where the first inequality uses the definition of $\rho(s)$ in Definition 3.1. The second inequality uses the fact that $\bar{\boldsymbol{\beta}}_j = 0$ when $j \in F^{(k)} - \bar{F}$. The third inequality is due to (20) and $\delta^{(k)} \geq \epsilon$. This proves the first desired bound of the lemma.

In order to prove the second and the third desired bounds of the lemma, we consider

$$\begin{aligned} \rho(s) \left\| \boldsymbol{\beta}^{(k)}_{F' - \bar{F}} \right\|_2^2 &\geq \rho(s) \left| F^{(k)} - \bar{F} \right| \delta^{(k)} \Big/ 2 \\ &\geq \frac{\rho(s)^2 \left| F^{(k)} - \bar{F} \right|}{2n \left| \left(F \cup F^{(k)} \right) - F^{(k-1)} \right|} \left\| X\boldsymbol{\beta}^{(k)} - X\boldsymbol{\beta}' \right\|_2^2 \\ &\geq \gamma^2 n^{-1} \left\| X\boldsymbol{\beta}^{(k)} - X\boldsymbol{\beta}' \right\|_2^2 \\ &\geq \gamma^2 \rho(s) \left\| \boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}' \right\|_2^2. \end{aligned} \tag{21}$$

The first inequality uses (20) and $F' - \bar{F} = F^{(k)} - \bar{F}$; the second inequality uses (19); the third inequality uses the assumption involving γ in the lemma; the fourth inequality uses the definition of $\rho(s)$ in Definition 3.1.

Therefore, we have from the last inequality above

$$\begin{aligned} \left\| \boldsymbol{\beta}^{(k)}_{F' - \bar{F}} \right\|_2 &\geq \gamma \left\| \boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}' \right\|_2 \\ &\geq \gamma \left\| \left(\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}' \right)_{F' - \bar{F}} \right\|_2 \\ &\geq \gamma \left\| \boldsymbol{\beta}^{(k)}_{F' - \bar{F}} \right\|_2 - \gamma \left\| \boldsymbol{\beta}'_{F' - \bar{F}} \right\|_2. \end{aligned}$$

By rearranging this inequality, we obtain

$$\begin{aligned} \left\| \boldsymbol{\beta}'_{F' - \bar{F}} \right\|_2 &\geq (1 - \gamma^{-1}) \left\| \boldsymbol{\beta}^{(k)}_{F' - \bar{F}} \right\|_2 \\ &\geq (\gamma - 1) \rho(s)^{-1/2} n^{-1/2} \left\| X\boldsymbol{\beta}^{(k)} - X\boldsymbol{\beta}' \right\|_2 \end{aligned} \tag{22}$$

where the second inequality uses (21). Therefore

$$\begin{aligned} n^{-1} \|X\bar{\boldsymbol{\beta}} - X\boldsymbol{\beta}'\|_2^2 \geq \rho(s) \|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}'\|_2^2 &\geq \rho(s) \|\boldsymbol{\beta}'_{F' - \bar{F}}\|_2^2 \\ &\geq (\gamma - 1)^2 n^{-1} \left\| X\boldsymbol{\beta}^{(k)} - X\boldsymbol{\beta}' \right\|_2^2 \end{aligned}$$

The first inequality uses the definition of $\rho(s)$ in Definition 3.1; the second inequality uses the fact that $\bar{\boldsymbol{\beta}}_j = 0$ when $j \in F' - \bar{F}$; the third inequality uses (22).

The above displayed inequality can be simplified to

$$(\gamma - 1)^{-1} \|X\bar{\boldsymbol{\beta}} - X\boldsymbol{\beta}'\|_2 \geq \left\| X\boldsymbol{\beta}^{(k)} - X\boldsymbol{\beta}' \right\|_2 \tag{23}$$

which implies that

$$\begin{aligned}&[1+(\gamma-1)^{-1}]\|X\bar{\boldsymbol{\beta}}-X\boldsymbol{\beta}'\|_2\\&\geq\|X\bar{\boldsymbol{\beta}}-X\boldsymbol{\beta}'\|_2+\left\|X\boldsymbol{\beta}^{(k)}-X\boldsymbol{\beta}'\right\|_2\\&\geq\left\|X\bar{\boldsymbol{\beta}}-X\boldsymbol{\beta}^{(k)}\right\|_2.\end{aligned}\tag{24}$$

Therefore, we obtain

$$\begin{aligned}&(\gamma-1)^2\left[(Q(\bar{\boldsymbol{\beta}})-Q(\boldsymbol{\beta}'))-\left(Q\left(\boldsymbol{\beta}^{(k)}\right)-Q(\boldsymbol{\beta}')\right)\right]\\&=(\gamma-1)^2n^{-1}\left[\|X\bar{\boldsymbol{\beta}}-X\boldsymbol{\beta}'\|_2^2-\|X\boldsymbol{\beta}^{(k)}-X\boldsymbol{\beta}'\|_2^2\right]\\&\geq(\gamma-1)^2n^{-1}\left[\|X\bar{\boldsymbol{\beta}}-X\boldsymbol{\beta}'\|_2^2-(\gamma-1)^{-2}\|X\bar{\boldsymbol{\beta}}-X\boldsymbol{\beta}'\|_2^2\right]\\&=[(\gamma-1)^2-1]n^{-1}\|X\bar{\boldsymbol{\beta}}-X\boldsymbol{\beta}'\|_2^2\\&\geq[(\gamma-1)^2-1][1+(\gamma-1)^{-1}]^{-2}n^{-1}\left\|X\bar{\boldsymbol{\beta}}-X\boldsymbol{\beta}^{(k)}\right\|_2^2\\&=(\gamma-1)^2[1-2\gamma^{-1}]n^{-1}\left\|X\bar{\boldsymbol{\beta}}-X\boldsymbol{\beta}^{(k)}\right\|_2^2.\end{aligned}$$

The first equality follows from the fact that $\boldsymbol{\beta}'$ is the least squares solution that minimizes $Q(\cdot)$ in F' and Proposition B.1. The first inequality uses (23). The second inequality uses (24).

By simplifying the above equation, we obtain the second desired bound of the lemma.

The third desired bound of the lemma follows from the following derivation:

$$\begin{aligned}Q(\bar{\boldsymbol{\beta}})-Q(\boldsymbol{\beta}')&=n^{-1}\|X\bar{\boldsymbol{\beta}}-X\boldsymbol{\beta}'\|_2^2\\&\geq\rho(s)\|\bar{\boldsymbol{\beta}}-\boldsymbol{\beta}'\|_2^2\geq\rho(s)\|\boldsymbol{\beta}'_{F'-\bar{F}}\|_2^2\\&\geq\rho(s)(1-\gamma^{-1})^2\left\|\boldsymbol{\beta}^{(k)}_{F'-\bar{F}}\right\|_2^2\\&\geq\rho(s)(1-\gamma^{-1})^2\left|F^{(k)}-\bar{F}\right|\epsilon\Big/2.\end{aligned}$$

The first equality follows from Proposition B.1, where we note that $\boldsymbol{\beta}'$ is the least squares solution that minimizes $Q(\cdot)$ in F' and $\mathrm{supp}(\bar{\boldsymbol{\beta}})\in F'$. The first inequality uses the definition of $\rho(s)$ in Definition 3.1. The second inequality uses the fact that $\bar{\boldsymbol{\beta}}_j=0$ when $j\in F'-\bar{F}$. The third inequality uses (22). The fourth inequality uses (20), $F'-\bar{F}=F^{(k)}-\bar{F}$, and the fact that $\epsilon\leq\delta^{(k)}$. This proves the third desired bound of the lemma. ■

The following result, which holds when FoBa terminates, is analogous to Lemma B.3.

Lemma B.4: Consider any $\bar{F}\subset\{1,\ldots,d\}$, and any $\bar{\boldsymbol{\beta}}\in R^d$ such that $\mathrm{supp}(\bar{\boldsymbol{\beta}})\subset\bar{F}$. At the end of FoBa algorithm in Fig. 4, let $s=|F^{(k)}\cup\bar{F}|$. We have

$$n^{-1}\left\|X(\bar{\boldsymbol{\beta}}-\boldsymbol{\beta}^{(k)})\right\|_2^2\geq\rho(s)\left|F^{(k)}-\bar{F}\right|\epsilon\Big/2.$$

In addition, we have either

$$Q(\boldsymbol{\beta}^{(k)})\leq Q(\bar{\boldsymbol{\beta}})-n^{-1}\left\|X\bar{\boldsymbol{\beta}}-X\boldsymbol{\beta}^{(k)}\right\|_2^2\Big/2$$

or

$$\begin{aligned}&\max\left[n^{-1}\left\|X\left(\bar{\boldsymbol{\beta}}-\boldsymbol{\beta}^{(k)}\right)\right\|_2^2,16\left[Q\left(\boldsymbol{\beta}^{(k)}\right)-Q(\bar{\boldsymbol{\beta}})\right]\right]\\&\leq16\epsilon\rho(s)^{-1}\left|\bar{F}-F^{(k)}\right|\\&\leq32\epsilon\rho(s)^{-1}\left|\left\{j\in\bar{F}:\bar{\boldsymbol{\beta}}_j^2<32\epsilon\rho(s)^{-2}\right\}\right|.\end{aligned}$$

Proof: Let $F'=\bar{F}\cup F^{(k)}$, and $\boldsymbol{\beta}'=\hat{\boldsymbol{\beta}}(F')$, we have from the optimality of $\boldsymbol{\beta}'$ in F' and Proposition B.1 that $Q(\boldsymbol{\beta}^{(k)})-Q(\boldsymbol{\beta}')=n^{-1}\|X(\boldsymbol{\beta}'-\boldsymbol{\beta}^{(k)})\|_2^2$. Therefore, similar to the derivation of (19), we can obtain by Lemma B.1 (apply the lemma with $\alpha=1$, $F=F^{(k)}$, and $\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}$) and the termination condition of FoBa (which requires that $\min_i\min_\alpha Q(\boldsymbol{\beta}^{(k)}+\alpha\mathbf{e}_i)\geq Q(\boldsymbol{\beta}^{(k)})-\epsilon$)

$$\begin{aligned}\epsilon&\geq\frac{\rho(s)}{\left|F'-F^{(k)}\right|}\left[Q\left(\boldsymbol{\beta}^{(k)}\right)-Q(\boldsymbol{\beta}')\right]\\&=\frac{\rho(s)}{n\left|F'-F^{(k)}\right|}\left\|X\boldsymbol{\beta}^{(k)}-X\boldsymbol{\beta}'\right\|_2^2.\end{aligned}\tag{25}$$

Moreover, since $\boldsymbol{\beta}^{(k)}$ is the solution when the backward step terminates in the immediate previous iteration, the backward step termination condition requires $Q(\boldsymbol{\beta}^{(k)}-\boldsymbol{\beta}_j^{(k)}\mathbf{e}_j)-Q(\boldsymbol{\beta}^{(k)})>0.5\delta^{(k)}\geq0.5\epsilon$ for all $j\in F^{(k)}$. Therefore, similar to the derivation of (20), we obtain from Lemma B.2 (apply the lemma with $F=F^{(k)}$, $\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}$, and $\boldsymbol{\beta}'=\bar{\boldsymbol{\beta}}$)

$$\frac{1}{\left|F^{(k)}-\bar{F}\right|}\left\|\boldsymbol{\beta}^{(k)}_{F^{(k)}-\bar{F}}\right\|_2^2\geq\epsilon/2.\tag{26}$$

Now the first desired bound of the Lemma follows from (26) using the same derivation as that of the first bound in Lemma B.3.

In the following, we consider two situations. In the first case, we assume that the following inequality holds:

$$\|X\bar{\boldsymbol{\beta}}-X\boldsymbol{\beta}'\|_2\geq3\left\|X\boldsymbol{\beta}^{(k)}-X\boldsymbol{\beta}'\right\|_2.\tag{27}$$

This inequality is the same as (23) with $\gamma=4$ in the proof of Lemma B.3. Using the same algebraic derivation below (23) leading to the second bound of Lemma B.3, we obtain

$$Q\left(\boldsymbol{\beta}^{(k)}\right)\leq Q(\bar{\boldsymbol{\beta}})-(1-2/4)n^{-1}\left\|X\bar{\boldsymbol{\beta}}-X\boldsymbol{\beta}^{(k)}\right\|_2^2.$$

This means that in this situation the second desired bound of the current Lemma holds.

In the second case, we assume that (27) does not hold. That is

$$\|X\bar{\boldsymbol{\beta}}-X\boldsymbol{\beta}'\|_2<3\left\|X\boldsymbol{\beta}^{(k)}-X\boldsymbol{\beta}'\right\|_2.$$

It follows that

$$\left\|X\bar{\boldsymbol{\beta}} - X\boldsymbol{\beta}^{(k)}\right\|_2^2 \le \left[\|X\bar{\boldsymbol{\beta}} - X\boldsymbol{\beta}'\|_2 + \left\|X\boldsymbol{\beta}^{(k)} - X\boldsymbol{\beta}'\right\|_2\right]^2 < 16\left\|X\boldsymbol{\beta}^{(k)} - X\boldsymbol{\beta}'\right\|_2^2 \le 16n\epsilon\rho(s)^{-1}\left|\bar{F} - F^{(k)}\right|. \quad (28)$$

The first inequality is the triangle inequality. The second inequality uses the assumption that (27) does not hold. The third inequality uses (25) and the fact that $F' - F^{(k)} = \bar{F} - F^{(k)}$.

Note that (25) also implies that

$$\left[Q\left(\boldsymbol{\beta}^{(k)}\right) - Q(\bar{\boldsymbol{\beta}})\right] \le \left[Q\left(\boldsymbol{\beta}^{(k)}\right) - Q(\boldsymbol{\beta}')\right] \le \epsilon\rho(s)^{-1}\left|\bar{F} - F^{(k)}\right|.$$

This result, combined with (28), leads to the first part of the third displayed inequality of the lemma.

Moreover, we have

$$\begin{aligned} &32\epsilon\rho(s)^{-1}\left|\left\{j \in \bar{F} - F^{(k)} : \bar{\boldsymbol{\beta}}_j^2 \ge 32\epsilon\rho(s)^{-2}\right\}\right| \\ &\quad \le \rho(s)\sum_{j \in \bar{F} - F^{(k)}} (\bar{\boldsymbol{\beta}}_j)^2 \le \rho(s)\left\|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(k)}\right\|_2^2 \\ &\quad \le n^{-1}\left\|X\left(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(k)}\right)\right\|_2^2 < 16\epsilon\rho(s)^{-1}\left|\bar{F} - F^{(k)}\right|. \end{aligned}$$

The first inequality is simple algebra. The second inequality uses $\boldsymbol{\beta}_j^{(k)} = 0$ when $j \in \bar{F} - F^{(k)}$. The third inequality uses the definition of $\rho(s)$ in Definition 3.1. The fourth inequality uses (28). Hence

$$2\left|\left\{j \in \bar{F} - F^{(k)} : \bar{\boldsymbol{\beta}}_j^2 \ge 32\epsilon\rho(s)^{-2}\right\}\right| \le \left|\bar{F} - F^{(k)}\right|$$

which implies that

$$2\left|\left\{j \in \bar{F} - F^{(k)} : \bar{\boldsymbol{\beta}}_j^2 < 32\epsilon\rho(s)^{-2}\right\}\right| \ge \left|\bar{F} - F^{(k)}\right|.$$

This implies the second part of the third displayed inequality of the lemma. ∎

Appendix C
Properties of Sub-Gaussian Noise

The following lemma is a standard empirical processes bound for sub-Gaussian random variables. The bound is used to derive probability estimates in our analysis.

Lemma C.1: Consider n independent random variables $\xi = [\xi_1, \ldots, \xi_n]$ such that $\mathbf{E}e^{t(\xi_i - \mathbf{E}\xi_i)} \le e^{\sigma^2 t^2/2}$ for all t and i. Consider vectors $\mathbf{g}_j = [g_{i,1}, \ldots, g_{i,n}] \in R^n$ for $j = 1, \ldots, m$, we have for all $\eta \in (0,1)$, with probability larger than $1 - \eta$

$$\sup_j |\mathbf{g}_j^\top(\xi - \mathbf{E}\xi)| \le a\sqrt{2\ln(2m/\eta)}$$

where $a = \sigma \sup_j \|\mathbf{g}_j\|_2$.

Proof: For a fixed j, we let $s_j = \mathbf{g}_j^\top(\xi - \mathbf{E}\xi) = \sum_{i=1}^n g_{i,j}(\xi_i - \mathbf{E}\xi_i)$; then by assumption, $\mathbf{E}(e^{ts_j} + e^{-ts_j}) \le 2e^{a^2t^2/2}$, which implies that for all $\epsilon > 0$: $\Pr(|s_j| \ge \epsilon)e^{t\epsilon} \le 2e^{a^2t^2/2}$. Now let $t = \epsilon/a^2$, we obtain

$$\Pr\left(\left|\mathbf{g}_j^\top(\xi - \mathbf{E}\xi)\right| \ge \epsilon\right) \le 2e^{-\epsilon^2/2a^2}.$$

This implies that

$$\Pr\left[\sup_j \left|\mathbf{g}_j^\top(\xi - \mathbf{E}\xi)\right| \ge \epsilon\right] \le m\sup_j \Pr\left[\left|\mathbf{g}_j^\top(\xi - \mathbf{E}\xi)\right| \ge \epsilon\right] \le 2me^{-\epsilon^2/(2a^2)}.$$

This implies the lemma. ∎

Next we would like to obtain a bound on the 2-norm between the estimated parameter and the true parameter. The proof requires the following simple covering number estimate taken from [22].

Proposition C.1: Consider the unit sphere $S^{k-1} = \{x : \|x\|_2 = 1\}$ in $\mathbb{R}^k$ $(k \ge 1)$. Given any $\varepsilon > 0$, there exists an ε-cover $Q \subset S^{k-1}$ such that $\min_{q \in Q}\|x - q\|_2 \le \varepsilon$ for all $\|x\|_2 = 1$, with $|Q| \le (1 + 2/\varepsilon)^k$.

The following result provides a concentration bound for any fixed k-dimensional projection of a sub-Gaussian random vector.

Lemma C.2: Let Assumption 3.1 hold. Let $\tilde{P}$ be any fixed $n \times n$ projection matrix of rank k. Then for all $\eta \in (0,1)$, with probability larger than $1 - \eta$

$$\|\tilde{P}(\mathbf{y} - \mathbf{E}\mathbf{y})\|_2^2 \le \sigma^2[7.4k + 2.7\ln(2/\eta)].$$

Proof: According to Proposition C.1, given $\epsilon_1 > 0$, there exists a finite set $Q = \{q_j\}$ with $|Q| \le (1 + 2/\epsilon_1)^k$ such that $\|\tilde{P}q_j\|_2 = 1$ for all j, and $\min_i \|\tilde{P}\boldsymbol{\beta} - \tilde{P}q_j\|_2 \le \epsilon_1$ for all $\|\tilde{P}\boldsymbol{\beta}\|_2 = 1$ (since $\tilde{P}\boldsymbol{\beta}$ is in a k dimensional space).

Lemma C.1 (with $m = |Q|$, and $a = \sigma\sup_j \|q_j^\top\tilde{P}\|_2 = \sigma$) implies that with probability $1 - \eta$

$$\sup_j \left|q_j^\top\tilde{P}(\mathbf{y} - \mathbf{E}\mathbf{y})\right|^2 \le 2\sigma^2\ln(2|Q|/\eta) \le \epsilon_2^2 = 2\sigma^2[k\ln(1 + 2/\epsilon_1) + \ln(2/\eta)].$$

Let $\boldsymbol{\beta} = \tilde{P}(\mathbf{y} - \mathbf{E}\mathbf{y})/\|\tilde{P}(\mathbf{y} - \mathbf{E}\mathbf{y})\|_2$, then there exists j such that $\|\tilde{P}\boldsymbol{\beta} - \tilde{P}q_j\|_2 \le \epsilon_1$. We have

$$\begin{aligned} \|\tilde{P}(\mathbf{y} - \mathbf{E}\mathbf{y})\|_2 &= \boldsymbol{\beta}^\top(\mathbf{y} - \mathbf{E}\mathbf{y}) \\ &\le \|\tilde{P}\boldsymbol{\beta} - \tilde{P}q_j\|_2\|\tilde{P}(\mathbf{y} - \mathbf{E}\mathbf{y})\|_2 + \left|q_j^\top\tilde{P}(\mathbf{y} - \mathbf{E}\mathbf{y})\right| \\ &\le \epsilon_1\|\tilde{P}(\mathbf{y} - \mathbf{E}\mathbf{y})\|_2 + \epsilon_2 \end{aligned}$$

where the first inequality is simple algebra, and the second inequality uses the definitions of ϵ_1 and ϵ_2. Now by rearranging the above inequality, we obtain

$$\|\tilde{P}(\mathbf{y} - \mathbf{E}\mathbf{y})\|_2 \le \epsilon_2/(1 - \epsilon_1) \le \sigma\sqrt{2[k\ln(1 + 2/\epsilon_1) + \ln(2/\eta)]}/(1 - \epsilon_1).$$

Let $\epsilon_1 = 2/15$, we obtain the desired bound. ∎

The following result uses Lemma C.2 to derive an oracle-inequality like bound that holds uniformly for all sparse estimators.

Lemma C.3: Given a fixed set $\bar{F} \subset \{1,\ldots,d\}$, with probability larger than $1-\eta$, we have for all $F \subset \{1,\ldots,d\}$ and all vectors $\bar{\boldsymbol{\beta}}$ such that $\mathrm{supp}(\bar{\boldsymbol{\beta}}) \subset \bar{F}$, the following statement holds:

$$\begin{aligned}&\|X\hat{\boldsymbol{\beta}}(F)-\mathbf{E}\mathbf{y}\|_2^2-\|X\bar{\boldsymbol{\beta}}-\mathbf{E}\mathbf{y}\|_2^2\\&\le n[Q(\hat{\boldsymbol{\beta}}(F))-Q(\bar{\boldsymbol{\beta}})]+2\|X\hat{\boldsymbol{\beta}}(F)-X\bar{\boldsymbol{\beta}}\|_2\\&\quad\cdot\sqrt{7.4|\bar{F}|+2.7|F-\bar{F}|\ln(16d)+2.7\ln(2e/\eta)}\sigma.\end{aligned}$$

Proof: For each $F \subset \{1,\ldots,d\}$, we define $\eta_{F\cup\bar{F}} = d^{-|F-\bar{F}|}\eta/e$. Since $F\cup\bar{F}$ is uniquely determined by the set $F-\bar{F}$ that contains $|F-\bar{F}|$ elements, there are at most C_d^j (d choose j) choices of $F\cup\bar{F}$ such that $|F-\bar{F}|=j$ ($j=0,1,\ldots,d$). It follows that

$$\sum_{F-\bar{F}\subset\{1,\ldots,d\}}\eta_{F\cup\bar{F}} \le \sum_{j\ge 0} C_d^j d^{-j}\eta/e \le \sum_{j\ge 0}\frac{d^j}{j!}d^{-j}\eta/e \le \eta.$$

Therefore, by taking union bound of Lemma C.2 for all $F \subset \{1,\ldots,d\}$, we obtain that with probability at least $1-\sum_{F-\bar{F}}\eta_{F\cup\bar{F}} \ge 1-\eta$: for all $F\subset\{1,\ldots,d\}$

$$\begin{aligned}&\|P_{F\cup\bar{F}}(\mathbf{y}-\mathbf{E}\mathbf{y})\|_2^2\\&\quad\le\sigma^2[7.4|F\cup\bar{F}|+2.7\ln(2/\eta_{F\cup\bar{F}})]\\&\quad=\sigma^2[7.4|F\cup\bar{F}|+2.7|F-\bar{F}|\ln d+2.7\ln(2e/\eta)]\\&\quad\le\sigma^2[7.4|\bar{F}|+2.7|F-\bar{F}|\ln(16d)+2.7\ln(2e/\eta)]\end{aligned}\quad(29)$$

where $P_{F\cup\bar{F}}$ is the projection matrix to the subspace spanned by columns $\mathbf{f}_j$ ($j\in F\cup\bar{F}$), and hence has rank at most $|F\cup\bar{F}|$. Note that the last inequality uses $7.4|F\cup\bar{F}| \le 7.4|\bar{F}|+2.7|F-\bar{F}|\ln 16$.

The above inequality implies that (see the equation shown at the bottom of the page). The first equality can be obtained from simple algebra. The second equality uses the fact that $\mathrm{supp}(\hat{\boldsymbol{\beta}}-\bar{\boldsymbol{\beta}}) \subset F\cup\bar{F}$, and thus $(\hat{\boldsymbol{\beta}}(F)-\bar{\boldsymbol{\beta}})^\top X^\top = (\hat{\boldsymbol{\beta}}(F)-\bar{\boldsymbol{\beta}})^\top X^\top P_{F\cup\bar{F}}$. The first inequality follows from the Cauchy-Schwarz inequality. The last inequality uses (29). ∎

The following result is similar to Lemma C.3 (which is useful for proving oracle inequalities), but it is more suitable for feature selection, where it is useful when $\|X\hat{\boldsymbol{\beta}}(\bar{F},\mathbf{E}\mathbf{y})-\mathbf{E}\mathbf{y}\|_2$ is small.

Lemma C.4: Given a fixed set $\bar{F}\subset\{1,\ldots,d\}$, with probability larger than $1-\eta$, we have for all $F\subset\{1,\ldots,d\}$ and all vector $\bar{\boldsymbol{\beta}}$ such that $\mathrm{supp}(\bar{\boldsymbol{\beta}})\subset\bar{F}$, the following two statements hold:

$$\begin{aligned}&\sqrt{n[Q(\hat{\boldsymbol{\beta}}(\bar{F}))-Q(\hat{\boldsymbol{\beta}}(F\cup\bar{F}))]}\\&\le\sqrt{\|X\hat{\boldsymbol{\beta}}(\bar{F},\mathbf{E}\mathbf{y})-\mathbf{E}\mathbf{y}\|_2^2-\|X\hat{\boldsymbol{\beta}}(F\cup\bar{F},\mathbf{E}\mathbf{y})-\mathbf{E}\mathbf{y}\|_2^2}\\&\quad+\sigma\sqrt{2.7|F-\bar{F}|\ln(16d)+2.7\ln(2e/\eta)}.\end{aligned}$$

Proof: We define $\eta_{F\cup\bar{F}}$ as in the proof of Lemma C.3. However, instead of using the rank-$|F\cup\bar{F}|$ projection matrix $P_{F\cup\bar{F}}$ to derive (29), we use the rank-$|F-\bar{F}|$ projection matrix $(P_{F\cup\bar{F}}-P_{\bar{F}})$ to obtain from Lemma C.2 that with probability $1-\eta$, we have for all $F\subset\{1,\ldots,d\}$

$$\begin{aligned}&\|(P_{F\cup\bar{F}}-P_{\bar{F}})(\mathbf{y}-\mathbf{E}\mathbf{y})\|_2\\&\quad\le\sigma\sqrt{7.4|F-\bar{F}|+2.7\ln(2/\eta_{F\cup\bar{F}})}\\&\quad\le\sigma\sqrt{2.7|F-\bar{F}|\ln(16d)+2.7\ln(2e/\eta)}.\end{aligned}$$

Moreover, using properties of projection operators and the closed form of least squares solution, we obtain

$$\begin{aligned}&\|(P_{F\cup\bar{F}}-P_{\bar{F}})(\mathbf{y}-\mathbf{E}\mathbf{y})\|_2\\&\ge\|(P_{F\cup\bar{F}}-P_{\bar{F}})\mathbf{y}\|_2-\|(P_{F\cup\bar{F}}-P_{\bar{F}})\mathbf{E}\mathbf{y}\|_2\\&=\sqrt{\|P_{F\cup\bar{F}}\mathbf{y}\|_2^2-\|P_{\bar{F}}\mathbf{y}\|_2^2}-\sqrt{\|P_{F\cup\bar{F}}\mathbf{E}\mathbf{y}\|_2^2-\|P_{\bar{F}}\mathbf{E}\mathbf{y}\|_2^2}\\&=\sqrt{\|(I-P_{\bar{F}})\mathbf{y}\|_2^2-\|(I-P_{F\cup\bar{F}})\mathbf{y}\|_2^2}\\&\quad-\sqrt{\|(I-P_{\bar{F}})\mathbf{E}\mathbf{y}\|_2^2-\|(I-P_{F\cup\bar{F}})\mathbf{E}\mathbf{y}\|_2^2}\\&=\sqrt{n[Q(\hat{\boldsymbol{\beta}}(\bar{F}))-Q(\hat{\boldsymbol{\beta}}(F\cup\bar{F}))]}\\&\quad-\sqrt{\|X\hat{\boldsymbol{\beta}}(\bar{F},\mathbf{E}\mathbf{y})-\mathbf{E}\mathbf{y}\|_2^2-\|X\hat{\boldsymbol{\beta}}(F\cup\bar{F},\mathbf{E}\mathbf{y})-\mathbf{E}\mathbf{y}\|_2^2}.\end{aligned}$$

In the above derivation, the first inequality is the triangle inequality. The first two equalities use properties of projection operators (Pythagorean equation). The third equality uses the fact that $P_F\mathbf{g}=X\hat{\boldsymbol{\beta}}(F,\mathbf{g})$ for all $F\subset\{1,\ldots,2\}$ and $\mathbf{g}\in R^n$ (which follows from the closed form solution of least squares problem).

Now, by comparing the previous two displayed inequalities, we obtain the lemma. ∎

$$\begin{aligned}&\|X\hat{\boldsymbol{\beta}}(F)-\mathbf{E}\mathbf{y}\|_2^2-\|X\bar{\boldsymbol{\beta}}-\mathbf{E}\mathbf{y}\|_2^2\\&=\|X\hat{\boldsymbol{\beta}}(F)-\mathbf{y}\|_2^2-\|X\bar{\boldsymbol{\beta}}-\mathbf{y}\|_2^2+2(X\hat{\boldsymbol{\beta}}(F)-X\bar{\boldsymbol{\beta}})^\top(\mathbf{y}-\mathbf{E}\mathbf{y})\\&=n[Q(\hat{\boldsymbol{\beta}}(F))-Q(\bar{\boldsymbol{\beta}})]+2(X\hat{\boldsymbol{\beta}}(F)-X\bar{\boldsymbol{\beta}})^\top P_{F\cup\bar{F}}(\mathbf{y}-\mathbf{E}\mathbf{y})\\&\le n[Q(\hat{\boldsymbol{\beta}}(F))-Q(\bar{\boldsymbol{\beta}})]+2\|X\hat{\boldsymbol{\beta}}(F)-X\bar{\boldsymbol{\beta}}\|_2\|P_{F\cup\bar{F}}(\mathbf{y}-\mathbf{E}\mathbf{y})\|_2\\&\le n[Q(\hat{\boldsymbol{\beta}}(F))-Q(\bar{\boldsymbol{\beta}})]+2\sigma\|X\hat{\boldsymbol{\beta}}(F)-X\bar{\boldsymbol{\beta}}\|_2\\&\quad\cdot\sqrt{7.4|\bar{F}|+2.7|F-\bar{F}|\ln(16d)+2.7\ln(2e/\eta)}\end{aligned}$$

The following lemma gives a bound on the infinity-norm of the difference between the estimated parameter $\hat{\boldsymbol{\beta}}(\bar{F})$ and the true parameter $\bar{\boldsymbol{\beta}}$ when the set of features $\bar{F}$ is known in advance.

Lemma C.5: Let Assumption 3.1 hold. Consider any fixed $\bar{F} \subset \{1,\ldots,d\}$ with $|\bar{F}| = \bar{k}$, and Let $\bar{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\bar{F}, \mathbf{E}\mathbf{y})$. For all $\eta \in (0,1)$, with probability larger than $1-\eta$, we have

$$\|\hat{\boldsymbol{\beta}}(\bar{F}) - \bar{\boldsymbol{\beta}}\|_\infty \leq \sigma\sqrt{(2\ln(2\bar{k}/\eta))/(n\rho(\bar{k}))}.$$

Proof: For simplicity, let $G \in R^{n\times\bar{k}}$ be the matrix with columns $\mathbf{f}_j$ for $j \in \bar{F}$. Let $\hat{\boldsymbol{\beta}}' \in R^{\bar{k}}$ and $\bar{\boldsymbol{\beta}}' \in R^{\bar{k}}$ be the restrictions of $\hat{\boldsymbol{\beta}}(\bar{F}) \in R^d$ to $\bar{F}$ and $\bar{\boldsymbol{\beta}} \in R^d$ to $\bar{F}$ respectively. By definition of $\hat{\boldsymbol{\beta}}$ as least squares solutions, and apply the closed form solution of least squares problems, we have $\hat{\boldsymbol{\beta}}' = (G^\top G)^{-1}G^\top\mathbf{y}$ and $\bar{\boldsymbol{\beta}}' = (G^\top G)^{-1}G^\top\mathbf{E}\mathbf{y}$. It follows that

$$\hat{\boldsymbol{\beta}}' - \bar{\boldsymbol{\beta}}' = (G^\top G)^{-1}G^\top(\mathbf{y} - \mathbf{E}\mathbf{y}).$$

Therefore, for $j = 1,\ldots,\bar{k}$

$$|\hat{\boldsymbol{\beta}}'_j - \bar{\boldsymbol{\beta}}'_j| = \left|\mathbf{e}_j^\top(G^\top G)^{-1}G^\top(\mathbf{y} - \mathbf{E}\mathbf{y})\right|.$$

Lemma C.1 (with $a^2 = \sup_j \|e_j(G^\top G)^{-1}G^\top\|_2^2\sigma^2$) implies that with probability larger than $1-\eta$, for all $j = 1,\ldots,\bar{k}$

$$\left|\mathbf{e}_j^\top(G^\top G)^{-1}G^\top(\mathbf{y} - \mathbf{E}\mathbf{y})\right| \leq \sigma\sup_j\left\|\mathbf{e}_j^\top(G^\top G)^{-1}G^\top\right\|_2\sqrt{2\ln(2\bar{k}/\eta)}.$$

According to Definition 3.1, $\rho(\bar{k})n$ is no larger than the smallest eigenvalue of $G^\top G$. Therefore, the desired inequality follows from the estimate

$$\left\|\mathbf{e}_j^\top(G^\top G)^{-1}G^\top\right\|_2^2 = \mathbf{e}_j^\top(G^\top G)^{-1}\mathbf{e}_j \leq 1/(n\rho(\bar{k})).$$

■

REFERENCES

[1] A. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 930–945, Mar. 1993.

[2] A. Barron, A. Cohen, W. Dahmen, and R. DeVore, "Approximation and learning by greedy algorithms," *Ann. Statist.*, vol. 36, pp. 64–94, 2008.

[3] P. Bickel, Y. Ritov, and A. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *Ann. Statist.*, vol. 37, no. 4, pp. 1705–1732, 2009.

[4] P. Bühlmann, "Boosting for high-dimensional linear models," *Ann. Statist.*, vol. 34, pp. 559–583, 2006.

[5] F. Bunea, A. Tsybakov, and M. H. Wegkamp, "Sparsity oracle inequalities for the Lasso," *Electron. J. Statist.*, vol. 1, pp. 169–194, 2007.

[6] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp, "Aggregation for Gaussian regression," *Ann. Statist.*, vol. 35, pp. 1674–1697, 2007.

[7] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[8] D. M. Chickering, "Optimal structure identification with greedy search," *J. Mach. Learn. Res.*, vol. 3, pp. 507–554, 2003.

[9] G. C. C. Huang and A. Barron, Risk of Penalized Least Squares, Greedy Selection and L_1 Penalization for Flexible Function Libraries, 2008, Yale Tech. Rep..

[10] C. Couvreur and Y. Bresler, "On the optimality of the backward greedy algorithm for the subset selection problem," *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 3, pp. 797–808, 2000.

[11] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.

[12] J. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari, "Composite objective mirror descent," presented at the COLT, 2010.

[13] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York: Springer, 2001.

[15] L. Jones, "A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training," *Ann. Statist.*, vol. 20, no. 1, pp. 608–613, 1992.

[16] V. Koltchinskii, "Sparsity in penalized empirical risk minimization," *Annales de l'Institut Henri Poincaré*, 2008.

[17] N. Littlestone, "Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm," *Mach. Learn.*, vol. 2, pp. 285–318, 1988.

[18] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[19] N. Meinshausen and P. Buhlmann, "High-dimensional graphs and variable selection with the Lasso," *The Ann. Statist.*, vol. 34, pp. 1436–1462, 2006.

[20] N. Meinshausen and B. Yu, "Lasso-type recovery of sparse representations for high-dimensional data," *Ann. Statist.*, vol. 37, pp. 246–270, 2009.

[21] S. Negahban, P. Ravikumar, M. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," presented at the NIPS, 2009.

[22] G. Pisier, *The Volume of Convex Bodies and Banach Space Geometry*. New York: Cambridge Univ. Press, 1989.

[23] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.

[24] L. Xiao, "Dual averaging methods for regularized stochastic learning and online optimization," *J. Mach. Learn. Res.*, vol. 11, pp. 2543–2596, 2010.

[25] C.-H. Zhang and J. Huang, Model-Selection Consistency of the Lasso in High-Dimensional Linear Regression, Rutgers Univ., 2006, Tech. Rep..

[26] T. Zhang, "On the consistency of feature selection using greedy least squares regression," *J. Mach. Learn. Res.*, vol. 10, pp. 555–568, 2009.

[27] T. Zhang, "Some sharp performance bounds for least squares regression with L_1 regularization," *Ann. Statist.*, vol. 37, no. 5A, pp. 2109–2144, 2009.

[28] T. Zhang, "Analysis of multi-stage convex relaxation for sparse regularization," *J. Mach. Learn. Res.*, vol. 11, pp. 1087–1107, 2010.

[29] P. Zhao and B. Yu, "On model selection consistency of Lasso," *J. Mach. Learn. Res.*, vol. 7, pp. 2541–2567, 2006.

Tong Zhang (M'10) received the B.A. degree in mathematics and computer science from Cornell University, Ithaca, NY, in 1994, and the Ph.D. degree in computer science from Stanford University, Stanford, CA, in 1998.

After graduation, he was with the IBM T.J. Watson Research Center in Yorktown Heights, NY, and Yahoo Research in New York. He is currently a Professor of statistics at Rutgers University, Piscataway, NJ. His research interests include machine learning, algorithms for statistical computation, their mathematical analysis, and applications.